An Explanation of Social Norms as Illustrated by Tipping

An honors thesis for the Department of Economics.

Benjamin Michael Glass

Tufts University, 2013

Table of Contents

Acknowledgments

In addition to the help of my thesis committee, I would like to acknowledge Jeffrey Luz-Alterman and Professor Gary Leupp for their help on Japanese and late medieval history. I also would like to thank Meghan Worsley for her helpful comments regarding altruistic motivations behind giving money to charity, as well as for her emotional support throughout the process of putting this paper together.

Section I: Introduction

Economics has had a rocky relationship with social norms, having never developed much of a standardized framework for modeling them. In fact, much of the investigation of social norms by economists boils down to guesswork; Kenneth Arrow famously conjectured that social norms are possibly "the reactions of society to compensate for market failures," and that social norms are mutually agreed upon to improve the economic efficiency of the system (Arrow 1971). Other economists have sought evolutionary explanations for social norms; they abandon the assumption of the materially self-interested rational actor and replacing it with the constraint of evolutionary fitness.[1]

But it is not entirely clear that social norms are even within the domain of economics per se. Many other social sciences have investigated the formation of social norms, including psychology, sociology, anthropology, and political science. For example, Elster (1989) argues against economists' attempts to understand social norms as arising from rational self-interest, collective self-interest, or evolutionary concerns. Rather, he suggests that emotional mechanisms that compel one to follow social norms are the answer, and that such a study fits in naturally with the understanding of human behavior promulgated by Emile Durkheim. An understanding of the effect of social norms on the allocation of resources in a society is critical to many sources of economic inquiry. Yet the development of these social norms is a subject matter that the discipline of economics has by and large not rigorously addressed outside of evolutionary models that lack significant predictive power.[2] Without a rigorous

[1] Examples include Frank (1987) and Frank (1988), as well as Nelson and Greene (2003).

[2] For example, while Frank (1988) develops a model for the development of evolutionary signaling mechanisms, he never can specify a resultant structural framework for utility functions that one could use as even the basis for doing empirical work. Similarly, Nelson and Greene (2003) do not even bother suggesting any kind of mathematical framework by which one could do empirical work.

understanding as to the forces behind the development of social norms, doing empirical work on phenomena beholden to these norms can become nearly impossible.[3]

The study of evolutionary processes almost tautologically must explain the development of social norms. Evolutionary processes select for the strategies that perform the best in the given environment, meaning that the problem of strategy selection for species can be expressed in terms of game theory, where reproduction is to be maximized rather than utility. The application of these strategies to modern situations should yield the behavior observed in contemporary humans. Furthermore, if the hunter-gatherer environment and biological limitations imposed on our ancient ancestors is sufficiently understood, it is plausible that we could derive new mathematical models of human behavior with greater sophistication and predictive power than the standard neoclassical model, if not the proposed replacements of Prospect theory, hyperbolic discounting, etc.

Yet the evolutionary approach to economics hits an intimidating obstacle in its implementation; after all, nobody is exactly sure of the conditions of humans before the advent of civilization, and the prospect of proposing statistical distributions for such considerations as food availability, probability of death in various activities, and systemic risk are factors of concern. For example, without some understanding of the distribution of covariances between variables of concern, it is not clear how to derive the ideal evolutionary strategy to deal with uncertainty.

Furthermore, the problem with biological limitations is immense. In principle, the problem is yet greater than the problem of mathematically modeling the human brain, since modeling the human brain is a prerequisite to modeling all human biological limitations. Yet the behavior of each neuron is determined by a system of nonlinear differential equations, and every neuron has complex interactions with other neurons in the system of billions of neurons that make up the human brain, let alone the

[3] In principle, Frank (1987) could be adapted to do empirical work, since it uses statistical distributions to model probabilities of betrayal. But the model was ostensibly not designed for empirical work, and thus lacks justifications for some of the assumptions. Furthermore, I am not aware of anybody who has tried.

entire nervous system.[4] In theory, an ability to compute the values of neuronal activity throughout the nervous system would definitively solve the problem; in practice, such an approach is little more than a pipe dream. While the evolutionary approach does not need to model human biology to that level of detail, it does need to have quantitative models for every biological process, as well as stochastic distributions for environmental concerns, to yield a model that produces quantitative predictions. This, too, appears to be a pipe dream.

Yet Frank (1988) uses the evolutionary model to explain the development of deviations from rationality and results from prisoner's dilemma games. He argues that emotions can act as pre-commitment mechanisms that allow for greater long-run chances of survival than would arise from behaving as a rational actor. This paper goes farther in connecting this insight with the "planner-doer" model described by Thaler and Shefrin (1981). The emotional mechanisms that arise in human behavior can be thought of as being determined by a planner who cares only about maximizing the long-run reproduction of one's genes. However, this planner is not necessarily completely rational, as his/her behavior was determined by evolutionary circumstances that are significantly different from those that are likely to arise in modern circumstances. Nonetheless, if one can understand how modern circumstances are translated into the circumstances likely to arise in prehistory, one can then begin to understand the behavior of the planner. For example, workplace competition can be interpreted as competition for being the successful hunter or gatherer, in which case status is the ultimate goal; money, or how much food one gets from one's success, is a proximate but secondary consideration.

In contrast, the doer is a rational actor who determines the behavior of the actor in any given situation, subject to a utility function that the planner has altered and limitations on behavior that the

[4] The basic mathematical model of a neuron is a system of four nonlinear differential equations, known as the Hodgkin-Huxley equations. See Hodgkin, Alan L., and Andrew F. Huxley. "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve." *The Journal of Physiology* 117.4 (1952): 500. More recent models are yet more complex, almost never yield an explicit solution, and are still only approximations of actual neuronal behavior. Finding computational solutions for vast networks of neurons is a challenge; finding an analytic solution is generally not considered a feasible goal for the foreseeable future.

planner has imposed. One of the most common examples is that most people would not even consider the possibility of murdering someone else for financial gain, even though the doer could maximize utility by doing so. This is because the planner has imposed a rule upon the doer that prohibits killing another person as an acceptable strategy in all but a few desperate circumstances. Alternatively, in the ultimatum game, the planner of the responder pre-commits well before the game is initiated to reject unfair offers by increasing the sense of anger or shame from accepting such an offer, thus changing the utility function of the doer to reject sufficiently unfair offers in cases like that of the ultimatum game.

The planner-doer model arises out of evolutionary situations in which cooperation is mutually beneficial for the two players, yet there is a large incentive to deceive. But nothing prevents this model from being applied to a wide variety of circumstances going way beyond those that motivated the existence of the model. People can commit themselves to redistribute resources, fight over trivially small potential gains, and refuse to make deals with those who slight them. It is in this manner that social norms are able to arise.

The goal of this paper is two-fold: to show how social norms arise and are sustained, and to explain how this model may be effectively applied. The benefit of the planner-doer model understood in the manner explained above is that most, if not all, behavior can be modeled with the techniques of neo-classical economics in the short run, although one must use more careful sociobiological techniques in the long run. Yet a model may well be considered plausible if it has an actor maximizing immediate welfare while maximizing what the planner interprets as the odds of reproduction in the long run.

In order to give structure to this model, and to focus on a subject matter that is of clear importance to the field of economics, I focus on the example of the payment of a voluntary gratuity in return for services rendered, also known as tipping. The phenomenon of tipping is of clear economic importance, being the source of tens of billions of dollars per year in income for those in the service industry (Azar 2004a). Yet the phenomenon exhibits characteristics that are difficult to explain with

models of rational behavior. Not only do people tip in places where the odds of having a repeated interaction with the same server are low, but the effect of repeated visits on tip size is negligible.[5] In contrast, Lynn (2006) reviews literature discussing the significant effects of race, alcohol consumption, and party size on the bill paid.[6] Yet Parrett (2006), while finding that men tend to tip more than women, did not find any effect of the method of payment on tip size. Meanwhile, Rind and Strohmetz (2001) find not only that good weather can increase the tip size, but even prove through a controlled experiment that *hearing* that there will be good weather causes an increase in the amount tipped.

So it appears that the determinants of the amount tipped have relatively little to do with rational considerations in any meaningful sense. Instead, it appears that Elster's argument of social norms as results of emotional processes dominates the explanation, suggesting that Lynn and Parrett are both correct in trying to understand tipping through psychological mechanisms.

Yet the biologically-determined planner-doer model does a relatively good job in explaining these phenomena as long as one keeps in mind the limitations of human rationality; furthermore, besides referring to the difference between more cognitive and emotional processes, the above explanations do not require a further referencing of psychological or biological processes.[7] The ease with which the model explains these phenomena is a testament to its elegance and general applicability. More importantly, by keeping in mind the planner-doer model, one can see how a failure of customers to understand to what degree improved service results from the change in how much they intended to tip can result in a higher social norm for tipping, which may make everyone better off. Furthermore, the

[5] Azar (2010) argues that there is no effect of repeated use of services on tip size for restaurants. While I think that it is a bit extreme to dismiss small increases in tip size that correlate with repeated use of service as white noise, the point of how small this increase tends to be is well-taken.

[6] While Lynn was reviewing literature related to restaurant tipping, Ayres, Vars, and Zakariya (2005) showed that racial discrimination also occurred when it came to tipping for cab service.

[7] Cognitive processes can be thought of as requiring effort to understand the situation. So the use of facts and figures to explain the situation can often make the process used to process the situation cognitive in nature. For example, Ariely (2009) shows that the more facts and figures are used in advertisements for charity (i.e., the more cognitive processes on the part of the viewer are activated, and thus compete with the emotional processes), the less often viewers give to that charity.

planner-doer model offers a compelling explanation for why some societies develop a social norm for tipping, but others do not. However, this explanation requires an understanding of a society's history and economic conditions, implying that both are important for the study of the tipping norm.

The complexity inherent in this biologically-determined planner-doer (BDPD) model makes it extraordinarily difficult to explicitly model in a general case scenario that would explain the various phenomena associated with tipping. Instead, the general characteristics of the model will be sketched out, with more mathematical material regarding the model relegated to the appendix. This should hopefully leave the main material accessible to an audience of non-economists.

When I refer to tipping, I will often use the terminology of restaurant tipping in the United States; however, the discussion will be applicable to all situations of tipping. Furthermore, the planner-doer model is structured in such a way that it can readily be applied to other subject matters as well. In particular, I discuss social norms relating to racism and marriage, among others.

Section II qualitatively describes the model in a general sense, illustrating how biological limitations on human behavior almost inevitably lead to signals being sent between humans, and how these signals can be used to partially circumvent problems in which there are Nash equilibriums that are less favorable than cooperative strategies, leading to the development of social norms. Some quantitative material is included towards the end of the section for illustrative purposes. Section III qualitatively describes how the model applies to tipping, and uses the model to explain some of the phenomena related to tipping listed above. The history of tipping will be briefly reviewed as well. Section IV provides a toy model of how utility functions can still be applied using the planner-doer model, although the structure of the utility function is left relatively flexible. The section describes how the social norm of tipping may evolve over time. Section V is analogous, but provides even less structure for the utility function, and thus has weaker results; its purpose is more illustrative of what is possible under the model. Section VI discusses the welfare implications of tipping as understood by the

planner-doer model; some emphasis is placed on the ambiguity of discussing welfare when referring to the planner-doer model. Section VII discusses the relevance of material in the previous sections and the appendix to a plethora of other social norms. Section VIII briefly concludes. The appendix contains more technical material that proves various useful results for the body of the paper, but these results are relegated to the appendix for expositional purposes.

For those interested in qualitatively understanding the phenomenon of tipping, sections II and III may be most useful; section VI may also be of some interest. For those who wish to understand the biologically-determined planner-doer model, sections II and VII may be of greatest relevance, although some material in the appendix may also be useful in this regard. For an understanding of how to utilize the planner-doer model in a formal model, sections IV or V would be useful. While sections II and III are referenced in all of the other sections, the other sections are designed to be intelligible independently of each other.

Section II: How Biological Limitations Lead to the Development of Social Norms

At first, one may imagine that the evolutionary process would select for rational actors. Such an argument is made in sources such as Sinn (2003), who argues that utility functions should maximize the opportunity to reproduce, and that such theoretical problems as introduced by phenomena like the Allais paradox need not be a concern simply because such results may hold in the lab, but will not in the field.[8] But this is not how evolution works. Yes, the behavior of human beings should tend to correspond to a function that selects the dominant strategies from those available in evolutionary circumstances, but those strategies readily include the sacrifice of rationality.

To see why evolution may select for deviations from rationality, first consider the case of purely rational actors in the case of the prisoner's dilemma.[9] In particular, imagine a tribe in which there are two dominant men in competition for the women of that tribe. Every time they come upon each other, they have a conundrum as to whether or not to attack each other. If one man killed the other, the remaining man would have little if any competition for the women of the tribe, and would be capable of having many children. Presumably the one who attacks first has the advantage, and may very well emerge unscathed. But if both attack, they may both be mortally wounded. The details of this problem could alter the exact set of dominant strategies, but the general setup suggests that each man might find it to be in his own interest to attack; he knows that failing to do so could easily mean death.

Yet each man has a flight or fight response hard-wired into him for this kind of situation.[10] In a situation of this level of danger, one's subconscious responses kick in to maximize the odds of survival. This is a particularly good strategy when faced with such an enemy as a tiger or a bear, where

[8] This is a point of no small contention. In short, while counterexamples to neoclassical economic theory can be readily found in the real world, John List (2003) and List (2004) show that at least some of these paradoxes diminish with the frequency with which actors make transactions, at least in that particular market. For an example of the limitations on the transfer of learned rationality, see Levitt, List, and Reiley (2010).
[9] This original need to circumvent the prisoner's dilemma was illustrated in Frank (1987) and Frank (1988).
[10] The details of this response can be found in, for example, Gray JA: The Psychology of Fear and Stress, 2nd ed. Cambridge, Cambridge, University Press, 1988. The response is triggered in the preganglionic neurons of the spinal cord, stimulating the sympathetic nervous system, as in Jansen et. al (1995).

instantaneous instinctual reflexes will serve someone far better than could conscious reasoning with Paleolithic technology. But in the situation of a fellow human as a combatant, this response gives away valuable information.[11] If the other man does not even see the flight-or-fight response kick in, he may figure out that his opponent is not interested in fighting. Or if one man looks pale and skittish while still not exhibiting aggressive signs indicative of the kind of adrenaline rush associated with an imminent attack, then the other man can figure out that his opponent is on his guard, but not about to strike. Given this information, the other man can design his strategy accordingly. So if one man looks sufficiently prepared for a conflict, but does not look overly aggressive, the other man might decide that he is not likely to be attacked, but would not likely benefit from attacking. In such a case, both men may benefit from being wary and ready, but not aggressive, as these dispositions will be apparent to each other and reinforce a mutually beneficial equilibrium that the two men would have preferred to combat in the first place, thus (partially) resolving the prisoner's dilemma.[12]

It is now possible that a pre-commitment to cooperation is an evolutionarily viable strategy. In particular, each man can evolve to be triggered by environmental and cognitive conditions to be less aggressive in the previously described situation. This could be due to the personal connection with the opponent, the tribal setting around his family and friends, or simply a hesitance to kill one's fellow man. In any case, being less aggressive in that situation will show, and the man will then be less likely to be attacked as long as he still looks wary. But note that this lack of aggression is a limitation on rationality. After all, while both men are more likely to find the peaceful equilibrium of the prisoner's dilemma with reduced aggression, each man in particular is less likely to pounce upon his opponent by reducing the necessary aggression that would make an attack more successful. This kind of cooperation would not be

[11] Note that while one consciously selects a strategy for generally dealing with various threats, the flight-or-fight response kicks in without conscious effort. Yet this response arises depending on the strategy selected.

[12] Note that this is not a complete resolution, as both men would most prefer not even having to worry about the situation; instead, both men are left being wary of each other. This is suggested as indicative of the limitations of the planner-doer model in resolving the prisoner's dilemma.

had between rational actors in such a situation as this, where a reputation cannot be built because there is no mechanism to punish bad behavior (since at least one of the men is then killed). Yet each man has an evolutionary incentive to not be rational in that situation, and thus cripples his own rationality in the name of survival.

It is important to note that this adaptation can only occur because a signal was already being sent before the adaptation. Without such a signal being sent out, the adaptation would not lead to cooperation; Frank (1988) elaborates on this point in detail. Yet it is obvious that humans indicate their sincerity through a variety of signals, many of which are to varying degrees beyond our control. Our emotional states have subconscious effects on our facial muscles, breathing, and skin color, and it can be difficult to consciously override these responses.[13] And even the process of overriding these responses can require emotional suppression, implying the firing of more cognitive centers of the brain, which itself changes behavior.[14] Furthermore, Gregg (2007) proved that lying takes more time than telling the truth, even when the lie will not be scrutinized, so even the attempted suppression of a signal can be a signal of behavior. In short, there is good reason to believe that people can pick up information on each other's intentions, and react accordingly.

To formalize this intuition, say that there are two players in a prisoner's dilemma. However, before the players make their moves in this game, they get to signal to each other. For either player, there is a probability that they can send this signal without making any pre-commitments at all. However, there is a chance that the player cannot send the signal without first compelling him/herself to follow a particular (irrational) course of action: if the player receives the same signal from the other player, than s/he will have to choose to cooperate. If we assume that players will always send signals if they can do so, and if we let p represent the probability that a player *cannot* send a false signal, then it

[13] Of course, some are exceptionally good at this, such as Adolf Hitler, who famously tricked Neville Chamberlin into thinking he had appeased Hitler, and Joseph Stalin who thought likewise a couple of years later.
[14] For example, see Carrión, Keenan, and Sebanz (2010).

will be worthwhile for a player to pre-commit to cooperation when they cannot send a false signal; the probability that both players cooperate is p^2, the probability that one player cooperates is $2p$, and the probability that neither player cooperates is $(1-p)^2$.[15]

One can generalize this situation to one of an iterated prisoner's dilemma with n trials (for $n \in \mathbb{N}$). Let each player know p with certainty, with p being the probability that a signal of willingness to cooperate until betrayed if the same signal is received from the other player is honest. So if both players signal to each other before playing the iterated prisoner's dilemma, then a player who honestly signals will cooperate the first round, and continue to cooperate up until and including the first round when the other player chooses betrayal, at which point the honest player will play betrayal for the remaining rounds. If there is a constant discount rate θ per round and both players signal cooperation with each other, then a player who sends a dishonest signal will betray the other player immediately or will cooperate until the last round. Furthermore, if both players are dishonest, then they will begin betraying each other at the exact same time. This suggests that cooperation for part of a prisoner's dilemma and betrayal in the middle should be due to either a varying discount rate θ or uncertainty regarding the probability of honest signaling p.[16]

A similar result holds in the *trust game*, which is like the prisoner's dilemma except that one player moves before the other one. The result is the same for rational actors; neither player cooperates. However, imagine that the player that moves second can pre-commit him/herself to cooperate if the other player cooperates first. Furthermore, imagine that doing so sends a signal to the first player that the second player cannot easily/likely send without first making this pre-commitment. In an analogous manner to what is described above, the player that moves first may then find it to be in his/her self-interest to choose to cooperate.

[15] See the appendix for a description of how one can parametrically test this model for both players cooperating with the same probability. However, note that this is the same as testing for symmetry in a Nash equilibrium, and does not speak to the truth of the model as a whole.
[16] This result is proven in the appendix.

In short, there are a plethora of situations in which people would like to pre-commit to various courses of action. While people can frequently make legally-binding contracts nowadays, it's hard to imagine that society could enforce these contracts before the development of government institutions. Even now, there are a plethora of situations in which people would like to pre-commit to one course of action, yet cannot create a contract that pre-commits them as they wish. For example, a monopolist would like to be pre-committed to setting prices very low if competition enters the market, but it has virtually no capacity to do so if the federal government enacts laws that prevent this behavior. So it makes sense that humans should be able to pre-commit themselves to behavior that is not materially self-interested in order to affect the behavior of others. Furthermore, it is relatively obvious that people do compel themselves to behave in an irrational manner, as attested to by the results of the ultimatum game.[17]

In general, there are a couple of ways that a person can avoid acting in a materially self-interested manner. One, s/he could compel him/herself to not consider certain options that would be in his/her material self-interest. Two, the person could alter his/her utility function to consider non-material concerns, such as moral issues. This is the exact model described by Thaler and Shefrin (1981), and they referred to their model as the planner-doer model. Their notion was that human behavior should be understood through the (metaphorical) interaction of a planner and a doer. The planner can establish rules limiting the behavior of the doer, and can also change the utility function of the doer. The doer then maximizes utility subject to the restrictions and changes of the planner.

However, the biological grounding of this model allows for even more intuition than the original planner-doer model. The model had previously conceived of human behavior as being decided by "a farsighted planner and a myopic doer." Instead, consider the planner to embody the evolutionary

[17] The ultimatum game is where one player offers a certain fraction of a potential amount of money, and the other player can either accept or reject the offer. If the offer is rejected, then neither player gets any money. A rational actor would accept any positive offer, but usually responders reject offers of a trivial amount of money.

interests of the actor in question. In the setting in which humans evolved, the planner is well-suited to maximize reproduction for the actor.[18] The doer can act rationally, but there is no reason why the doer cannot make inter-temporal choices or even take the future behavior of the planner into account. Thus, the doer can choose behavior that compels the planner to put restrictions on the future behavior of the doer.

There is some evidence to support this model. Most obviously, people frequently choose to cooperate in situations in which game theory predicts defection, even when the games are so simple as to make bounded rationality an implausible explanation. In addition, the model suggests that one's ability to falsely signal cooperation to other players is determined by biological (especially neurological) conditions, although a shared cultural background with the other player may facilitate communication. In accord with this model, Segal and Hershberger (1999) showed that cooperation in the prisoner's dilemma is more likely when the players are more genetically similar. Similarly, Cesarini et al. (2008) showed that genetics significantly influences how cooperative either player is in the trust game.

But as George Box famously said in Box and Draper (1987, p. 474), "Essentially, all models are wrong, but some are useful." The benefits of this model arise from more than its explanatory power. It suggests that issues can be looked at from an evolutionary perspective, and that people will be able to pre-commit themselves in a manner that is to their evolutionary advantage. Furthermore, social norms can now be explained. Those in a position of power in a community can promote others to internalize social norms that are beneficial for the community, and thus for those in power. They can do so by promoting the norm to those with whom they interact, and those people continuing to promote the norm to others in the community. In addition, those with power can compel those without power to internalize values of loyalty to the current system, even though such values may pre-commit the disenfranchised to behavior that is not in their self-interest.

[18] The exact function that the planner is trying to maximize is briefly discussed in section VI, but this consideration is not directly relevant to this paper.

But once social norms are established, they can remain powerful and unmoving even as material conditions change. In particular, norms that are more rigidly defined can persist indefinitely even as everybody continues to follow them only out of a desire to conform.[19] For example, the position of the emperor in Japan has survived thousands of years, and reverence for the emperor has been generally considered a noble value for all of modern Japanese history. Yet the emperor has been virtually powerless for most of the past millennium. The notion of reverence for the emperor was established and cultivated by the imperial family when it held power, and the norm of imperial reverence has persisted despite the changing economic and political conditions of the country.

The model also suggests that the ability of society to develop social norms that improve societal welfare is limited. In a war, the soldiers of both sides would benefit from agreeing to not kill each other. However, such an agreement would require signaling between soldiers, and this is usually impossible.[20] Instead, soldiers from each side signal to other soldiers on their own side, and these soldiers also receive signals from those higher up in the chain of command. Since neither side is signaling cooperation to the other, no agreement can be reached to diminish the violence.

In general, social norms require signaling between human beings to take hold. This is why those higher up in a social hierarchy have such an inordinate influence on the development of social norms; they can exert a lot of "pressure" on those beneath them in the hierarchy to embrace the established norms, while those below them rarely can effectively organize enough to resist unwanted norms.[21] One person cannot see the signals being sent from a large group of people, making it very hard to

[19] This point is demonstrated in the appendix. Note that the desire to conform is instilled by the planner in the BDPD model.
[20] A notable exception is the famous Christmas Truce on the western front of World War I in 1914.
[21] By pressure, I am referring to incentives that some can provide others to embrace various social norms.

communicate signals of pre-commitment to certain norms or strategies without a social hierarchy to propagate these signals.[22]

Finally, this model yields a good explanation for cognitive dissonance.[23] When the planner (partially) pre-commits to a value, yet the doer successfully manages to act in discord with that value, then the discomfort felt by the actor may be the punishment that the doer bears for circumventing the planner's previous restriction. For example, the planner may find it worthwhile to weakly commit the doer to honor previous agreements in business dealings, but the planner does not want the doer to pass up too large of an opportunity to benefit from a dishonest business dealing. So when the opportunity arises for the doer to benefit significantly from such an opportunity, the doer suffers the pain of cognitive dissonance, but still manages to pursue the ideal behavior as far as the planner is concerned. Note that the utility of the doer is reduced due to cognitive dissonance, but the pain from cognitive dissonance had no direct effect on the goal of the planner to maximize evolutionary fitness. Whereas the doer is primarily concerned with utility, the planner is primarily concerned with reproduction.

It could be that the planner cannot possibly elaborate on the ideal behavior of the doer for each possible situation. Instead, the planner could commit the doer to relatively vague principles, and later fill in the details as various situations arise. The more that the course of action seems to clash with the notions to which the doer has already been pre-committed, the more cognitive dissonance the doer feels. Someone may feel very strongly that murder is wrong, but fuzzy situations can arise with euthanasia, self-defense, abortion, and choosing between lives to save.[24]

[22] Of course, there are exceptions. For example, labor unions work because unions create a social hierarchy to help organize workers and propagate norms that demonize scabbing during strikes.

[23] Note that cognitive dissonance has been a part of economic analysis for decades. For example, Akerlof and Dickens (1982) examined the effect of cognitive dissonance on the adoption of safety measures in the workplace. However, there has been very little (if any) work done on formally modeling the phenomenon of cognitive dissonance. While this paper also does not formally model cognitive dissonance, this section and the appendix suggest ways in which such modeling could be done.

[24] This point is elaborated on in the appendix.

In conclusion, the biologically-determined planner-doer model posits that human behavior is ultimately determined by a planner that restricts the behavior of the doer in order to maximize fitness in some manner previously determined by evolutionary forces. This implies that cooperative results can arise when standard game theory predicts betrayal, while vengeance may motivate behavior to the disadvantage of everyone involved on the basis of principle. Furthermore, social norms can arise due to the ability of those in a position of power to promote the norms to those beneath them. These values could promote the interests of everybody in society, and they usually promote the interests of those in power. However, they can also logically follow from other values embraced by the society; values that can be persuasively argued to be implied by other values will tend to compel some interest due to a desire to avoid cognitive dissonance. The implication is that social norms are determined from the material and historical conditions from which they arise.

Section III: Explanatory Power of the Model for the Social Norm of Tipping

Azar and Tobol (2008) estimates the amount of money left as a tip in the United States to amount to $44 billion per year in the food industry alone. But Azar (2010) and Lynn (2006) consistently found that the frequency with which one visits a restaurant has no effect on the amount tipped. The phenomenon of tipping appears to have a substantial effect on the economy, yet it stubbornly refuses to yield even in part to standard neo-classical analysis; the model of the rational self-interest actor cannot explain tipping without reputational concerns. Azar (2004b) argued that the norm of tipping developed as a method of resolving the trust game. However, Parrett (2006) shows evidence that tipping is motivated by both concerns for reciprocity (rewarding good service) and letdown aversion (conforming to expectations). While concern for reciprocity helps to resolve the trust game, letdown aversion constitutes nothing more than a redistribution of wealth from the customer to the server (and indirectly to the restaurant). Furthermore, these motivations do not explain the variance in the strength of the norm of tipping between various societies. Harris, Lynn, and Zinkhan (1993) demonstrate strong correlations between certain characteristics a society might have and the strength of the norm of tipping, but they cannot give more than a cursory explanation of the outliers.

However, applying the biologically-determined planner-doer (BDPD) model allows for an explanation of the results found in Harris, Lynn, and Zinkhan (1993) while also explaining many of the results that others have found on the effects of environmental and social factors on amounts tipped. The BDPD model has substantial explanatory power, but requires an understanding of the history of a society in order to be fully employed; in fact, a study of the history of a society is completely necessary to understand the strength of the norm of tipping in that society today. An examination of western societies in comparison with Japanese society reveals why the norm of tipping is strong in the United States, weak in much of Europe, and virtually non-existent in Japan. Furthermore, the BDPD model suggests that the norm of tipping would be very powerful in Japan if it had ever developed, but that the

norm has never gotten started due to the way in which Japan developed from a feudal society. While the United States has many similar characteristics as Japan does as a society, the U.S. never had a feudal society, and thus does not still embrace societal values that are preventing the norm of tipping from getting started in Japan.

Azar (2004b) illustrates how the interaction between a server and a customer in the service industry can be perceived as a trust game. The server would like to provide the minimal service required unless sufficiently financially compensated for extra effort. The customer would be willing to pay for the extra service. However, the service is provided first, and the customer later decides how to allocate financial resources between him/herself and the server. If it is in the mutual interest of the server and the customer to have payment exchanged for effort, it follows that tipping may be the cooperative result of the biologically-determined planner-doer (BDPD) model. The customer sends the server a signal that good service will be rewarded, the server tends to provide better service out of a rational desire to be rewarded, and the server is (usually) rewarded with more than enough of a tip to compensate the server for the extra service provided.

However, there are a couple of major caveats to this description of the phenomenon of tipping. First of all, the BDPD model requires the relevant strategies to already be understood. If the phenomenon of tipping is not already established, then the customer cannot signal his/her willingness to reward good service with a tip. Second of all, there is no guarantee that the BDPD model could not describe societal pressure being exerted on customers to compel them to tip (or tip more) when they do not otherwise benefit from the action. In fact, it appears quite clear that people tip at least in part to conform to the values of society. The BDPD model suggests that it is necessary to examine the history of a society in order to determine whether a norm developed out of the interest of those involved to find a mutually beneficial alternative to self-interest, or if the norm developed out of the interest of one group with influence compelling another to embrace values that benefit the first group at the expense of the

second. The history of tipping in the West suggests that the social norm of tipping arose from a more complicated situation than a simple trust game.

Segrave (1998, p. 1-2) argues that tipping arose from a custom of the late middle ages in which servants would receive money from the guests of their masters after serving these guests during their stay.[25] Similarly, there are widespread accounts of money being used to buy good service in the 18th century, with payment for good service in the restaurant industry among these accounts.[26] Cheyney (1900) points out that serfdom in England was all but dead by the 16th century; while the English maintain an aristocracy to this day, that aristocracy has lived in a market-based economy for over four hundred years. Furthermore, the lords had mostly volunteered to give their serfs their freedom; it's possible that it was more profitable to employ workers than to own them at the time. The English aristocracy chose to embrace capitalism, and tipping was one of their first adaptations.

Schein, Jablonski, and Wohlfahrt (1984, p. 20) and Haley (2011, p. 172) say that the custom of tipping was exported to the United States in the early 19th century, although it did not become widespread amongst the upper class until after the Civil War according to Haley (2011, p. 172). Yet the upper class had a choice from only a handful of high-class restaurants, and thus tipping for them was a rational behavior; their reputations depended on tipping well. For the upper class, there was no need for the planner to step in to resolve the trust game between customer and server; the concern for reputation should have already played that role for wealthy members of "society" who only dined at the highest-quality restaurants.

[25] Segrave (1998) described this practice, known as the giving of vails, as having arisen during Tudor England (1485-1603). While the sources referenced are not to my liking in terms of quality, the claim is plausible due to the more widespread use of currency in English society by that point.

[26] There is a competing theory described by Segrave (1998, p. 5-6) that tipping arose from a concept of giving "drink money" as a reward for good service. For example, the word "tip" may be derived from the word "tipple", meaning "drink". A third theory, suggested by Segrave and endorsed by Schein, Jablonski, and Wohlfahrt (1984, p. 19) is that tipping arose in coffee shops where empty cans with the message "To Insure Promptitude" or "To Insure Promptness" were filled with coins by customers eager for fast service. Furthermore, Schein, Jablonski, and Wohlfahrt (1984, p.20) pushes the flourishing of tipping in the United States forward to the turn of the 20th century. However, Schein, Jablonski, and Wohlfahrt (1984) does not use footnotes and was written earlier; I consider it to be a less trustworthy source in general.

However, the middle class did not have this same incentive to maintain the practice of tipping, and the middle class of the United States was fiercely against the custom. Segrave (1998, p. 54-5) describes those in the middle class lambasting the acceptance of tips as debasing to everyone involved; there were widespread reports of the opulent wealth of the servers who accepted tips even after all hope had died out of outlawing or otherwise getting rid of the practice of tipping. Schein, Jablonski, and Wohlfahrt (1984, p.21) point out the rise of the Anti-Tipping Society of American in the mid-1900s, and notes that subsequently laws were passed in Mississippi, Georgia, South Carolina, Tennessee, Arkansas, Washington and Iowa that outlawed tipping. However, the Supreme Court ruled that such laws were unconstitutional in 1919. The last such significant efforts to kill the practice of tipping died out in the 1910s in Great Britain according to Segrave (1998, p. 25-27) and a decade later in the United States according to Haley (2011, p. 180-85).

However, the middle class continued to resent the norm while the upper class propped it up. Segrave (1998, p. 66) describes magazines criticizing the practice of tipping throughout the 1940s, while admitting that the practice was here to stay. It was the upper class that continued to prop up the norm. In the 1920s, the nouveau riche and Hollywood stars ostentatiously tipped to show off their wealth. The Great Depression hurt the practice of tipping, but afterwards scarce labor and resurgent wealth combined to strengthen tipping as never before; Schein, Jablonski, and Wohlfahrt (1984, p.21) report that even private clubs had to allow for the tipping of bartenders and servers in order to attract necessary labor. Segrave (1998, p. 66) documents members of the upper class being put on the put on blacklists for being lousy tippers even into the 1940s.

Schein, Jablonski, and Wohlfahrt (1984, p.22) describe the social norm of tipping as being significantly weakened in Europe over the course of the century due to the proliferation of the service charge. However, this has not completely stifled the tipping norm. To give a few examples, while tipping is considered abhorrent in Iceland, Great Britain generally has a 10 percent tipping norm when a

service charge is not included, tipping is not expected but allowed in France when there is a service charge, and in Hungary tips are expected even when there is a service charge according to Schein, Jablonski, and Wohlfahrt (1984, p.142-4).

The picture painted above regarding the development of the tipping norm is not one of a resolution to the trust game. Instead, it is largely about a battle between social norms and a struggle over status. The upper class sought to flaunt their wealth while ensuring themselves good service, as in Haley (2011). Meanwhile, the middle class described tipping as extortionist for the customer and humiliating for the server; this argument is illustrated (and still partially advocated) in Schein, Jablonski, and Wohlfahrt (1984, p. 22).

It makes sense that those in the middle class would dislike the incentive that tipping gave servers to provide better service to those with more wealth; however, Segrave (1998, p. 22) points out that all efforts by those of the middle class to boycott the practice in restaurants were complete failures. Haley (2011, p. 181-6) points out that some effort was made to provide cafeteria-style service in order to avoid tipping altogether, but these institutions never acquired the status of sit-down restaurants. Thus, while the middle class may have generally despised the norm of tipping, the individual member within the class found it to be in his/her self-interest to continue to participate in the practice.

This is completely in line with the BDPD model described in the previous section. While Arrow (1971) postulated that social norms developed in order to improve societal welfare, it is unclear by his reasoning why a norm to refuse to patronize restaurants that supported tipping would not develop in the middle class of the United States; after all, such a norm would compel many restaurants to abandon the practice, which would be beneficial for the entirety of the middle class at the time.[27] However, the BDPD model describes social norms as arising from interactions between individuals and propagated by those with positions of power in society. Those with power in the late 19th century United States were

[27] You could argue that this happened in Europe to an extent, but the practice of tipping was not extinguished on the continent even with the increasing use of service charges.

almost always wealthy, and without powerful people propagating such a norm throughout the middle class as boycotting restaurants that condone tipping, those within the middle class could not effectively signal to each other their pre-commitment to this practice, and thus could not develop the norm.

But the BDPD model also relies on the importance of cognitive dissonance; this effect explains why different societal values interplay with the strength of the norm of tipping. For example, Harris, Lynn and Zinkhan (1993) point out that tipping is more prevalent in societies that value economic achievement over social relations (which they describe as "masculine values"), while it is less prevalent in societies that value more social relations (which they describe as "feminine values"). They also showed that societies that are more tolerant of stark economic inequality and power differences tend to have stronger norms for tipping.[28] It makes sense that those who believe in the "American dream", that hard work will be rewarded, would feel cognitive dissonance upon refusing to financially reward good service. Furthermore, if one were opposed to those with more money getting better service, then one might refuse to participate in the system of restaurant tipping based on principle. In fact, Haley (2011, p. 180) points out that the middle class portrayed tipping as "a risk to American democracy" in an analogous manner to the risk posed by the meddling of moneyed interests in American politics. That there is much less discontent with tipping nowadays may be indicative that the middle class is currently more tolerant of more extreme economic inequality.

Thus, the BDPD model implies that Parrett (2006) is correct in noting that people tip both to reward good service and to conform to an established norm. The first motivation arises from the signals that the server and the customer send back and forth to each other, leading to the customer pre-committing to reward good service with extra payment. The second motivation arises from the battles

[28] They also pointed out that societies that are less tolerant of uncertainty tend to have stronger tipping norms. However, it is not clear that this results from cognitive dissonance. Workers may wish to have more control over their pay, and customers over their service quality, when people in that society are more opposed to uncertainty. However, it is unclear whether the people in this society are altogether risk adverse or if they are pre-committed to a norm that compels them to avoid risk and remain in control. If it is the former, then this is not cognitive dissonance, but merely a difference in utility functions.

over societal values waged by different groups in society. The outcome of these conflicts arose from both the relative influence of the groups at odds and the generally agreed-upon societal values that acted as both tools and limitations for the dialogue.

Furthermore, it is not reasonable that anybody embraced tipping merely for the status boost that it may have provided them. One, tipping is not a particularly ostentatious display of wealth to anyone besides the server (and possibly those in one's own party). Two, the practice of tipping would already have to exist in order for it to become valued as a status symbol. In order to signal to others one's high status, the signal must first already be well-established. As argued in Frank (1988), biological signals can only be sent purposefully if they would already be observable without intentionally sending the signal. Thus, frogs with deep croaks attract mates because a frog's croak tended to be deeper when it was larger, and thus frogs began to develop deeper croaks in order to attract mates. Similarly, one cannot tip for status unless one who is of high status would tend to tip more often anyways. Tipping for better service provides the natural explanation for how tipping got started; afterwards, people could have tipped for status.

From the BDPD model, we can understand tipping as originally arising as a mechanism purposefully designed to resolve the trust game. Once the practice began to develop, various societal values helped to strengthen or smother the norm. In the United States, the middle class generally resisted tipping due to the improved service that the wealthy could now afford, and thus the middle class was being denied. However, the accepted societal values (and legal system) of the United States prevented all efforts to resist the development of the tipping norm. In Europe, the situation was more mixed, with many restaurants imposing a service charge in order to weaken the norm and encourage customer service in that manner. The significantly weaker tipping norm in much of Europe is easily explained by the difference in societal values between Europe and the United States.

But the societal values of a society can only partially explain the strength of the social norm of tipping. Japan has all of the characteristics that Harris, Lynn, and Zinkhan (1993) identify as positively correlated with the strength of the social norm of tipping. However, the norm of tipping in Japan is so weak as to barely exist.[29] Cho (2006) similarly attempted to determine what cultural difference between Japan and the United States could explain the difference in strength of the tipping norm; yet he found all such explanations lacking, instead proclaiming that the tipping norm was so weak in Japan simply because "Japanese do not expect to be tipped." As Schein, Jablonski, and Wohlfahrt (1984, p. 146-7) says,

> Tipping in Japan presents some problems because it is generally frowned upon. Custom and a sense of delicacy cause individuals to be offended by having money handed to them directly. If a tip is to be given, it should be placed in an envelope or put in a position where others will not see it. Sometimes a tip is left on a tray or it is arranged to be given in a discreet manner. In Japan, tipping is not customary in most day-to-day procedures. Tipping causes Japanese people to feel like beggars. So, in most circumstances it is a good idea not to tip. Instead you might offer a small gift as a token of appreciation.

The social norm of tipping never developed in Japan, but its failure to be developed had nothing to do with cognitive dissonance. Instead, the values of the aristocracy in Japan led to the proliferation of an uncomfortable attitude with money that prevents the social norm of tipping from developing.

[29] Although not as stark of a case, New Zealand also acts as an outlier in that it has the characteristics correlated with a strong tipping norm, but does not have such a strong norm at all. While I do not examine New Zealand in detail, perhaps it is similar to Japan in that its history created values that make it difficult for the country to develop a norm for tipping. It is worth noting that there are no countries that are outliers in that they have a notably stronger norm for tipping than their characteristics would suggest.

Until the mid-19th century, Japan was a somewhat feudal society. Throughout the 1500s, a couple of hundred warlords known as daimyo split up Japan and fought each other for power.[30] However, after the Battle of Sekigahara (1600), the regime was de facto united under the rule of the Tokugawa Shogunate.[31] The daimyo of Japan were all vassals under the shogun from the time of that battle until the fall of the shogunate in 1868. Furthermore, Earl (1964) describes the shogunate as immediately adopting the traditional Confucian notion that the merchants were the lowest class in Japanese society, whereas the samurai class (including the daimyo) was highest in status. Hane (1991, p. 163-4) says the vassals owed the shogunate "loyalty and prescribed obligations", although the shogun still left them some autonomy. The shogunate spent the first half of the 17th century swallowing up the domains of 120 daimyo families, but largely left the remaining domains alone thereafter. Instead, the shogunate instituted the policy of *sankin kotai*, or alternate attendance, in which each daimyo would spend every other year in the capital of Edo. This policy protected the shogunate from the daimyo by weakening the connection that the daimyo had with their domains, but also by draining the financial resources of the daimyo through the financing of two separate homes and the maintenance of an expensive entourage; since the daimyo could not accumulate funds, they were not in a position to challenge the shogunate until the system was abolished in the mid-19th century.

While the sankin kotai system drained the funds of the daimyo, the samurai found themselves in increasingly difficult financial straits because they largely depended on the stipends provided by their lords. Meanwhile, Hane (1991, p. 224-6) describes the development of capitalism in the cities of Japan causing the samurai and daimyo to have a higher, more expensive standard of living. Yet the merchants

[30] Hane (1991, p. 114-6) describes the daimyo as having arisen from a power vacuum caused by the Onin War (1467-77). However, the term *daimyo* can be used to refer to the warlords who ruled before the onset of these wars; yet these rulers were qualitatively different, since their authority arose in principle from the shogunate. In contrast, the daimyo of the 16th century had a much more local base of power.

[31] This is a stark simplification, and does not take into account the previous unification efforts of Oda Nobunaga or Toyotomi Hideyoshi, nor of the remaining loyalties to Toyotomi Hidetada until 1615. Furthermore, in theory the emperor is supposed to be in charge. Hane (1991) makes a good account of these details; however, they are beyond the scope of this paper.

were increasingly prospering with the development of the new economic system. This created discontent between the samurai and merchants classes. For example, Yodoya Tatsugoro was infamous for accumulating a massive fortune due to the multitude of daimyo that were indebted to him. In order to end the embarrassing state of affairs, the shogunate confiscated Yodoya's fortune in 1705 on the grounds that he was living in an extravagant fashion that was beyond the limits that his status as a merchant imposed.

When Commodore Perry opened Japan up to trade with the United States, he unleashed a lot of this discontent, ultimately causing the fall of the shogunate. Many of the people who would end up wielding political power in Japan were followers of Yoshida Shoin, an ultra-nationalist who was arrested for attempting to stow aboard Commodore Perry's ship and visit the United States in order to "get to know his enemy." He was a disciple of Sakuma Zozan, who insisted that Japan copy the scientific proficiency of the West while maintaining the morality of the East, i.e., traditional Japanese morality. This was the sentiment followed by many of those who later held power in Japan.[32]

So Japan entered the modern world with elite that remained dedicated to the moral values of the samurai aristocracy of the shogunate. Greenfeld (2001) argues that those samurai that led the Meiji Restoration sought for the samurai to participate in business enterprises for the benefit of the nation of Japan; they argued that the merchant class would not protect the interests of the nation. Business owners invoked the national good to justify pay cuts, and even nowadays the Japanese have an aversion to achieving business success without a sense of loyalty to one's company. From this historical background, it is no surprise that money is perceived in a negative light in Japan, and the transaction of money has a more sullying character to it than it does in the United States.

[32] Again, this is from Hane (1993, p. 256-8), but it is a dramatic simplification of the thinking of those who fought to overturn the shogunate and those who ran Japan in the late 19th century. For a more complete description, see Greenfeld (2001).

From this historical background, the BDPD model can now explain why Japan did not develop a strong tipping norm while the United States did. To see this, note that when one person acts altruistically for another, and that other person unexpectedly attempts to pay for the altruistic behavior, the first person will frequently become enraged or offended due to the offered payment. To see why this is the case, note that the second person is sending a signal to the first person that the behavior was not altruistic, but materialistically motivated by the ensuing payment. The offense taken by the first person compels them to reject the offered payment. This is clearly not in the material interests of the first person.

But if the first person is acting altruistically because that person is pre-committed to a social norm, then that person would suffer from cognitive dissonance by accepting the payment. It is a logical contradiction that the first person both acted in the interests of the second person out of benevolence *and* accepted payment for that behavior. Even if the first person could convince him/herself that his/her behavior was altruistic, but the second person was paying by mistake, it would still violate the generally accepted norm of not tricking people into giving you their money if the first person then took the second person's money.

However, note that the damage is done by accepting *any* payment for the altruistic behavior; the size of the payment is irrelevant. To see why, note that the second person clearly feels compelled to pay in return for the altruistic behavior. The amount paid is then a valuation of that service. To see why it must be a valuation of service quality, note that the BDPD model suggests that the social norm of payment after altruistic behavior is a solution to a trust game. But the service quality will be minimal unless the social norm incentivizes service quality by encouraging the beneficiaries of that service to pay more for better service. Therefore, any payment that follows altruistic behavior sends a signal that the other person must reject in order to avoid the pain of cognitive dissonance.

Thus, the BDPD model implies that anybody who acts in an altruistic manner (without intending to be compensated) will dislike being paid based on principle, but will be happier with the payment as the payment becomes larger.[33] If the norm of tipping is not well-established in a society, the server will either provide the minimal service that the manager requires (and can enforce) or will altruistically provide better service without expectation of compensation. In that situation, the customer would actually *punish* the server by leaving a tip unless that tip were large enough to compensate for the offense of tipping.

The importance of the history of Japan is that the Japanese aristocracy has successfully promulgated norms that increase one's status as one is able to show less concern with money. Since being overly concerned with money is perceived negatively in Japan, accepting payment for service reflects even more poorly on the server when the service was provided without intention of receiving payment. This is why Schein, Jablonski, and Wohlfahrt (1984) recommends giving a gift instead of a tip if possible; the gift incentivizes good service while not dealing in money, thus circumventing (at least some of) the disapprobation surrounding the exchange of money for such service.

Since Japan has the characteristics that Harris, Lynn, and Zinkhan (1993) suggest would tend to result in a strong tipping norm, it stands to reason that the tipping norm *would* be strong in Japan if it ever got off of the ground. But the societal values of Japan make it difficult for the norm to get off the ground. In this sense, the norm of tipping seems to be *stuck at zero* in Japan, meaning that it would develop and thrive if it somehow already existed, but currently the expected tip amount is stuck at zero. This is in contrast with much of Europe, where the tipping norm exists but is weak, and the United States, where the tipping norm thrives.

[33] For example, I have informally polled some friends of mine regarding the following question: if you have sex with somebody, would you prefer to be paid $100, $20, or $0? They consistently respond that they most prefer getting paid $0, but then prefer being paid $100 to $20.

The BDPD model does a strikingly good job qualitatively describing the strength of the tipping norm in various countries. But the model also implies that the amount tipped should have certain characteristics. First of all, the amount tipped should go down as there are increasing cultural barriers between the customer and the server. Note that the BDPD model relies on signals being sent between the two people; these signals are biological and psychological in nature, but rely on a shared cultural understanding in order to effectively find and interpret these signals. It stands to reason that religious or racial differences would tend to lead to smaller tips on average, and that furthermore these smaller tips would arise from different understandings of the tipping norm. In addition, gender *could* make a difference, while method of payment should not.

There is overwhelming evidence to support these claims. Lynn and Katz (2011) found that people who regularly attended religious services tipped worse than the general public; furthermore, the amount tipped varied less with service quality with those who regularly attend church than those who do not. This is consistent with the notion that religious people are having difficulty effectively signaling with their (presumably non-religious or differently-religious) server ; the amount that religious people tip have a lot to do with conforming to the norm of tipping, and has little to do with a resolution to the trust game that Azar (2004b) imagines. Ayres and Zakaria (2005) found that black customers tipped less and black drivers got tipped less in the taxicab industry. Lynn (2006) reviews the literature that indicates that racial differences significantly reduce tip amounts, and that these reductions have to do with black customers not understanding the customary amount to tip. Lynn (2004) shows that black people tend to tip a constant amount (independent of bill size) more often than white people. Parrett (2006) shows that the method of payment is unrelated to the amount tipped, but men tip more than women do.

Finally, note that the BDPD model relies on psychological mechanisms to pre-commit the customer to reward good service. This implies that environmental conditions could change the amount tipped. In fact, this is demonstrated in a couple of papers. Rind and Bordia (1996) showed that writing a

smiley face on the check led to a higher tip for waitresses (but lower for men). While it is a famous result that nicer weather leads to better tips, Rind and Strohmetz (2001) found in a controlled experiment that even *hearing* that the weather would be good led to increased tip amounts regardless of how the weather turned out to be. Lynn (2006), Azar (2010), and Parrett (2006) show a correlation between service quality and tip amount, as the BDPD model predicts would happen. However, this correlation is weak, since other psychological factors can throw off the customer's subjective valuation of the service quality. While the cognitive faculties of the customer may evaluate the service quality, it's the emotional faculties of the customer that compel the customer to tip.

The BDPD model does a remarkable job describing the prevalence and characteristics of the tipping norm. In addition, the model implies that certain results that cannot be empirically tested must be true. First of all, certain countries such as Japan can get stuck at zero, meaning that they have effectively no norm for tipping. Yet these countries would develop a strong norm for tipping if they already had such a norm. The welfare implications of this result are briefly discussed in section VI.

Second of all, people frequently tip to reward good service *as if* they were buying the service, although environmental influences can significantly disturb this phenomenon. The implication is that a market for good service is created without the traditional market mechanisms usually necessary to allow for the existence of such a market. In addition, if the customers perceive their money to buy them more service than it actually does, it follows that customers will tip more than they otherwise would. In particular, this means that people will tip more if they confuse the amount of improved service that they receive with the increased amount that they *intend* to tip, instead of the increase in the average amount tipped. The welfare implications of this result are also discussed in section VI.

Section IV lays out the formal model demonstrating these two results. Section V generalizes by avoiding the representative consumer. The only point that does not carry over into section V is that one cannot tell if ignorance of the improved service resulting from the stronger tipping norm causes the

tipping norm to become stronger. If customers are sufficiently pessimistic about the effect that their intended tip amount has on service quality, then it could be that the amount that people tip fluctuates and does not synchronize, and everybody weighs the reduced quality of service more than the improved quality of service, leading to everybody presuming that their intended tip amounts are relatively ineffective at improving service quality.

Section IV: Formally Modeling Tipping Using the Representative Customer

I start off by constructing the model customer. Note that the customer's utility function will have more than merely material concerns; as the previous section described, the utility function of the actor is maximized by the doer, but is shaped by the planner. The planner makes the actor more concerned about conforming to society, being fair and/or kind to the server, and other such matters. The customer's utility function should also be continuous, since a small change in any particular variable should not radically change the customer's utility. The utility function can be described as:

$$U(r, m, s, k) \tag{1}$$

In the above, *r* is the deviation of the amount tipped from some social norm, *m* is the amount of money with which the customer must part, *s* is the quality of service (as experienced by the customer), and *k* represents the amount of kindness shown to the server.[34] All of these variables will be constructed such that they cannot take on negative values. The idea is to describe the utility of the customer as a result of social, monetary, and intrinsic motivations in addition to service quality. I can define *r* and *m* as follows:

$$r = |C_1(s) + C_2(s)x - t| \tag{2}$$

$$m = x + t \tag{3}$$

In the above, $C_1$and C_2 are constants, *t* is the amount tipped, and *x* is the size of the bill. Note that $t \in [0, \infty)$; the tip must be nonnegative. Equation (1) has the following properties:

$$\frac{d}{dr}U(r, m, s, k) \leq 0 \tag{4}$$

$$\frac{\partial}{\partial r}U(0, m, s, k) = 0 \tag{5}$$

$$\frac{\partial}{\partial m}U(r, m, s, k) < 0 \tag{6}$$

[34] The kindness variable can encompass other motivations such as guilt, embarrassment, and even euphoric impulses. The term is merely meant to represent the effects of intrinsic motivations that one might assume to be independent of societal influence.

$$\frac{\partial}{\partial s}U(r,m,s,k) > 0 \quad (7)$$

$$\frac{\partial}{\partial k}U(r,m,s,k) > 0 \quad (8)$$

The value of *s* is a strictly increasing function of *t* through mechanisms that will be explained in a bit. However, one may very well describe *k* as a function of *t* as follows:

$$\frac{d}{dt}k(t) \geq 0 \quad (9)$$

$$k(0) = 0 \quad (10)$$

In other words, the model asserts that the utility function of the customer is a function of to what degree one has fulfilled one's social obligations *r*, how much money one has lost *m*, how good the service was *s*, and how kind the customer is being to the server *k*. The rest of the equations pick up some relatively obvious assumptions and identities, such as utility goes up as the quality of service goes up, and the amount of money spent is the sum of the size of the bill and the amount tipped. Equations (2) and (5) could use explaining. Equation (2) essentially asserts that there is a linear equation relating the size of the tip with the bill size, with the quality of service determining the coefficients. Furthermore, this and equation (4) assert that the further away the tip is from this "socially ideal" tip, the more utility goes down. Equation (5) ensures that the utility function is differentiable for all values of *t*.

In addition, we need a limiting assumption to rule out ridiculous behavior. It is implausible that somebody would find it to be in his/her interests to give as large of a tip as possible; it must be the case that the intrinsic desire to be kind to a server that has been kind to you must diminish over time. Let's say that, given all variables not dependent on *t*:

$$\exists T > 0, \forall t \geq T, \ \frac{dU}{dk}\frac{dk}{dt}(t) + \frac{dU}{dr}\frac{dr}{dt}(t) + \frac{dU}{dm}\frac{dm}{dt}(t) + \frac{dU}{ds}\frac{ds}{dt}(t) \leq 0 \quad (11)$$

One can simplify this slightly to say that *ceteris paribus*:

$$\exists T > 0, \forall t \geq T, \ \frac{dU}{dt} = \frac{dU}{dk}\frac{dk}{dt} + \frac{dU}{dr}\frac{dr}{dt} + \frac{dU}{dm} + \frac{dU}{ds}\frac{ds}{dt} \leq 0 \quad (12)$$

From these assumptions, one gets the following lemma:

Lemma 1: *From the above assumptions, given all variables not dependent on t, for some T as described in equation (12), there is a value of* $t \in [0, T]$ *such that utility is maximized at that point.*

This lemma is the formalization of the observation that, given the fact that the marginal intrinsic benefit of tipping doesn't remain too strong, somebody would find it to be in his/her interests to tip a finite amount. This point is somewhat trivial, and the proof is in the appendix.

To model the behavior of the server, I say that the server is able to gather some information on how much a customer may intend to tip through subconscious body language, tone of voice, etc.[35] A larger intended tip implies a larger potential tip, increasing the incentive for good service. In addition, a larger average tip allows for a larger potential tip, similarly incentivizing good service. Thus, letting $\bar{t}$ represent the (nonnegative) average amount tipped, we can express the quality of service as follows:

$$s(\bar{t}, t) \tag{13}$$

$$\frac{\partial s}{\partial \bar{t}} > 0 \qquad \forall \bar{t} \in [0, \infty) \tag{14}$$

$$\frac{\partial s}{\partial t} > 0 \qquad \forall t \in [0, \infty) \tag{15}$$

Note that I am claiming that, once a tip is intended, the customer becomes compelled to tip that amount. In reality, one commits oneself through one's disposition to tip a certain amount independent of whether an actual amount intended to be tipped is ever decided upon. For the purpose of simplicity, we need only acknowledge that the amount tipped is predetermined to a point by one's disposition during the meal.

[35] The server may also use such information as race, sex, clothing, and with whom the paying customer(s) is/are dining. How the server acquires this information is irrelevant, as long as the customer affects these signs through increasing his/her intended tip amount. For example, a customer who intends to tip handsomely may be on a date, wear expensive clothing, or simply get along well with the server. There is a problem in that other information may not be under even limited control by the customer; the most prominent example is race. However, this is a deviation from the model of the representative consumer, and is addressed in the next section.

As established in lemma 1, there is some value, say t^*, such that the utility function is maximized at that value given all other variables not dependent on the amount tipped, as well as the average amount tipped. Furthermore, let's assume for the purpose of simplicity that utility is maximized at only one point, so a consumer should not be indifferent to two amounts that s/he could tip. However, there is good reason to believe that the consumer lacks the information necessary to maximize utility. In fact, even given a perfect understanding of his/her utility function, s/he would lack a complete understanding of the equation for the level of service experienced, as expressed in (13). Instead it will be conducive to model a learning process by which the consumer approaches an understanding of expression (13).

Let there be an infinite number of periods indexed by *n*, and say that s_n is a function of $\bar{t}$ and *t* such that s_n becomes a closer and closer approximation of expression (13) over time.[36] This represents the process by which customers learn the effect of how much they tip on service quality. What results is the following theorem:

Theorem 1: Given the above setup and given $\bar{t}$*, let* $t_n = argmax\_t\ U\big(r(t), m(t), s_n(\bar{t}, t), k(t)\big)$*.Then* $\lim_{n\to\infty} t_n = t^*$.

In other words, the customer will get arbitrarily close to selecting the profit-maximizing tip over time. The proof is in the appendix.

So far, the model makes some fairly tame claim that with a sufficiently effective learning mechanism, one can learn how much to dispose oneself to tip over time. But now suppose that the customer instead assumes that the service is only a function of their own tip amount, and not on the average amount tipped. In other words, they conflate the correlation between their tip and service

[36] Formally, let $\{s_n\} \to s$ uniformly. A brief justification of this formalization can be found in appendix A2. Also, note that the consumer is assumed to care only about the utility from this visit. Formally, the discount rate for the utility of the next period is set to zero.

quality with causation between these two phenomena. Imagine an infinite-period game in which everybody has the same utility function and reacts to the same information in the same manner. Let:

$$S(t) = s(t, t) \tag{16}$$

Note that the customer's tip has replaced the average tip. Also, let $\{S_n\}$ be a set of functions of solely the amount that the customer tips, such that S_n becomes a closer and closer approximation of expression (16) over time.[37] In other words, everybody is learning as if the average tip did not matter. While the specifics of this learning process are elaborated on in the formal model, it is easy enough to understand when every consumer is identical; as consumers tip more, the service improves both due to the larger individual tip and the larger average tip, causing the consumer to think that s/he has a larger individual effect on service quality than in reality. The same logic applies as consumers tip less, implying that the derivative of equation (16) is larger than the derivative of the original social function with respect to the individual amount tipped.

This is a relatively straightforward way of thinking about the process of norm growth through learning, but is meant to be only an example as to how such learning could occur. Given all variables not dependent on the tip given (which does *not* include the average tip this time) people seek to maximize the following function:

$$P(t) = U(r(t), m(t), S(t), k(t)) \tag{17}$$

Note that if we denote the average tip in that period by $\overline{t_n}$ and the customer's tip in that period by t_n, then $\overline{t_n} = t_n$ in every period because everybody has the same utility function and starts off tipping the same amount. Let t^I represent the ideal amount to tip when maximizing this function. Finally, let

$$P_n(t) = U(r(t), m(t), S_n(t), k(t)) \tag{18}$$

[37] Formally, let $\{S_n\} \to S$ uniformly. Again, a brief justification of this formalization can be found in appendix A2.

This is the function that customers will be maximizing in period *n*. What results is the following theorem:

Theorem 2: Given the above setup, let $t_n = argmax\_t\ P_n(t)$*.Then* $\lim_{n\to\infty} t_n = t^I$*. Also,* $t^I > t^*$*.*

Thus, $\exists N$ *such that* $\forall n > N$*,* $t_n > t^*$*.*

This theorem says that the amount tipped will be greater in every time period when the customers are under this misconception than without it; furthermore, the amount that they tip while under this misconception will eventually be larger than the amount that they would tip to maximize utility if they understood the effect of the average tip on how good the service they receive is. The proof of this theorem is in the appendix, but the idea is that people will try to figure out what effect their propensity to tip has on service quality, and will overestimate it's effect because they are not aware of the effect of the changing social norm, or average tip, on the quality of service.

This line of thinking is more realistic than assuming that consumers are completely aware of the determinants of service quality. Even though customers may be quite aware that a stronger social norm of tipping, and thus a high average amount tipped, leads to better service (to which people who have visited Europe can attest), the customer does not perceive fluctuations in the average amount tipped nearly as quickly as does the server. In addition, customers are attempting to gauge how good the service was that they received through various means, but usually not by going through a checklist of what the server did right and wrong. Instead, customers tend to rely on their moods and feelings to indicate the quality of service.[38]

Of course, this line of reasoning quickly brakes down when we abandon the notion of the representative model consumer; in other words, once we allow for the possibility (in fact the certainty) that consumers vary in their exact preferences and initial tip amounts, then it becomes far from certain that every consumer will eventually tip more due to the learning process described above. For example,

[38] For example, there is a rich literature on the social psychology behind the effect of good weather on the amount tipped. In fact, studies including Rind and Strohmetz (2001) show that simply informing customers that the weather will be good *in the future* increases the amount tipped.

it could well be the case that there are initially two groups of people, the first of which initially tips much less than the second group. Over time, the first group could find it to be quite implausible that such small tips on their parts could motivate service much, while the second group finds it quite likely that their large tips do motivate good service. Suppose that the second group is a good deal larger than the first, so that the average amount tipped goes up over time. Then these observations could be encouraged by the fact that both groups may see service improve, reinforcing each group's own conclusions. Thus, the first group would tip less and less over time, while the second group would instead tip more and more. Certainly the first group of people would tip much more if they effectively took into account the fact that the average amount tipped is being significantly bolstered by the second group, as then the first group would realize that their own tips had a greater effect on service quality than previously thought, and they simply didn't notice because the service quality was being partially bolstered due to the second group.

Furthermore, one could imagine that the tendency for the learning process to increase the amount that customers tip would grow stronger if one included societal pressure to conform to the average amount tipped. For instance, in the above example it might instead be the case that both groups feel bashful about tipping such a divergent amount from what the average tip is, and thus each group tips a more moderate amount, causing their learning process to be similar to what it would be like if both groups tipped the same amount. Thus, one might figure it to be more likely that the learning process will cause both groups to tip more than they would have otherwise.

These results are beyond the confines of this present setup; the subsequent section is dedicated to the exposition of such a technical model, and the subsequent proofs of the related theorems. It is the purpose of the subsequent section to demonstrate the persistent power of the ideas developed in this section when most of the remaining unreasonable assumptions are dropped. Yet these ideas still make sense without the rigorous mathematical modeling to demonstrate their intellectual coherency. The

societal pressure to conform, in addition to the learning process by which one figures the relationship between tip amount and service quality, should tend to lead to people tipping more than they should otherwise; in particular, these two processes should lead to people tipping more on average than they otherwise would.

One consideration upon which we can touch is that a sufficiently small tip will likely be perceived by the server as insulting, creating an incentive for the customer to rather tip nothing than tip too small an amount. To model this consideration, reconsider the kindness function as follows:

$$\frac{\partial k}{\partial t} \geq 0 \qquad \forall t > 0 \tag{19}$$

$$k(0) = 0 \tag{20}$$

$$k(t_n) < 0 \qquad t \in \left(0, \beta(\overline{t_{n-1}})\right) \tag{21}$$

$$\frac{d}{dt}\beta(t) < 0 \qquad \forall t > 0 \tag{22}$$

$$k(t_k) - k(u_k) = k(t_l) - k(u_l) \qquad \forall k, l \in \mathbb{N}, t_k, t_l, u_k, u_l > 0 \tag{23}$$

Note that equation (23) merely ensures that the tip that maximizes utility does not change due to a changing kindness function over time.[39] This system of equations is not significantly different from what was previously described about the kindness function, but equations (19) through (21) and the mean value theorem together imply that the kindness function is discontinuous when the tip is zero. This makes intuitive sense; giving a very small tip is often considered to be less kind than giving no tip at all, as a small tip is seen as insulting. Finally, we want to assume that:

$$s(0, t) = const. \tag{24}$$

That is to say, servers do not look for the signs of a good tipper when the average is zero.

[39] There is some abuse of notation, as technically the kindness function should have a period subscript, as the kindness function changes over time. However, in appendix A2 I show that equation (23) implies that the kindness function has the same derivative in every period for any positive tip, so we can ignore this detail for our purposes.

From this setup, it seems like an initial tipping norm of zero could be particularly stable. This intuition is formalized in the following theorem, which technically describes the contrapositive of this sentiment:

Theorem 3 (Sticky Zero Theorem): *Given the above model of the representative consumer, with the kindness function modeled by equations (19) through (23),given all other variables there is an amount to tip that maximizes the utility function in any period. If the initial amount that the customer tips is positive, then the amount that the customer tips will eventually surpass the amount that would maximize the individual's utility function given the average amount tipped if that amount is positive, just as described in theorem 2. However, if the initial amount that the customer tips is zero, then the customer may continue to not tip at all indefinitely.*

This theorem points out that the amount that consumers tip can get stuck at the value zero; i.e. once customers stop tipping, they may not start tipping again even if they would be better off with a stronger tipping norm. After all, once the norm of tipping dies off, servers may find it insulting to receive a modest tip, in a similar manner to how one may be put off by someone offering to buy one's place in line for the bus for five dollars. For most people, such an amount should be worth the extra minute of their time waiting to get on the bus (and perhaps mildly worse pick of a seat on the bus). Yet few people would be willing to accept such a deal.

In line with the recent proliferation of work on path-dependence in economic growth and international trade, it's worth noting that a similar result arises from this model. In particular, note that the history of what has been is critical to understanding how the norm of tipping is developing. For example, an alien who were to take a look at the United States and Japan at the present time might presume that both countries should have the same social norm for tipping because both have very similar economic conditions that may be similarly conducive towards tipping, as discussed in Harris, Lynn, and Zinkhan (1993). However, the alien would not realize that Japan has previously been stuck at

zero, i.e. they have no tipping norm of which to speak, and thus are still unable to form a tipping norm. In contrast, the United States has had over a hundred years to develop its tipping norm, and thus by this point has developed one of the most powerful tipping norms in the developed world.

Finally, several prominent scholars from various disciplines who have investigated the norm of tipping have speculated that the norm is particularly bad for societal welfare. Of course, the very idea of societal welfare becomes muddy when one begins to allow for the possibility of non-materialistic concerns for consumers, as one must to allow for tipping without repeated visits. That being said, this model suggests that it need not be the case that societal welfare declines due to tipping. After all, if employers were hurt by the norm of tipping, then they would resist it by instead imposing a service charge. Instead, they choose to utilize the norm of tipping to incentivize their workers, as described in Azar (2004b).[40] Similarly, employees would likely find other jobs if they found working in restaurants that employed a tipping norm to be to their detriment to a degree that employers could not compensate. In other words, it ought to be the case that tipping is good for both employers and employees on principle, as both parties are actively choosing to participate in this system.[41]

Furthermore, note that the customers are similarly maximizing utility. In fact, some careful consideration will reveal that the utility of the representative customer is maximized under the learning model described in theorem 2. So for any positive tip as the norm, customers should be working through the process of tipping to maximize their utility, thus further increasing their welfare. Thus, it appears that everybody wins.

In fact, it looks as if the only case in which customers suffer is when the norm gets stuck at zero. When tipping doesn't happen at all, then customers cannot slowly increase the norm for tipping, as any

[40] In fairness, employers may also be saving money through avoiding payroll taxes, and also by allowing their employees to do the same and recouping some of the employees' increased earnings. Such tax fraud is likely common, but exact figures are elusive due to the illegal nature of the activity.
[41] There will be more on this subject matter in section VI.

positive tip of such a small magnitude would insult the server, and not lead to significantly improved service.

The important takeaways are the overarching importance of history in the development of a social norm such as tipping and the effect that limited rationality can have on service quality. In the next section, I attempt to generalize the above model to get beyond the representative consumer. However, one finds that one cannot have results of the same power as in this section without the assumption of the representative consumer. It may be that a completely different formalization is necessary to capture the intuition of this section without using the model of the representative consumer.

Section V: Formally Modeling the Development of Tipping Through Learning

This is a generalization of section IV in which the representative consumer model is abandoned. The BDPD model is still implicitly used, in that the utility function contains elements imposed by the planner, while the doer is presumed to maximize the utility function. Note that, just as in the previous section, the behavior of the server is considered completely determined by the amount tipped by the customers. Thus, we only need look at the utility function and information of the consumers.

I begin by imposing a structure to the utility function that is sufficiently flexible to allow for nearly any explanation for the forces that drive the norm of tipping. Although I have detailed structure that will not end up being used in the construction of these proofs in more than a superficial manner, this format is provided for the benefit of those who would like to do empirical work with this structure. Suppose that there is a closed economy with *N* customers, with each customer tipping some amount *t*. Furthermore, we can define $\bar{t}$ as the unordered N-tuple $\bar{t} = \{t_1, t_2, \dots, t_N\}$ and say that the customers have continuous utility functions of the form:[42]

$$U(\bar{t}, t) = f\big(r(\bar{t}, t), m(t), s(\bar{t}, t), k(s, t, \bar{t})\big) \quad (1)^{43}$$

In the above, $f(w, x, y, z)$ is differentiable in all of its arguments. In particular:

$$\frac{d}{dr} f(r, m, s, k) \leq 0 \quad (2)$$

$$\frac{d}{dr} f(0, m, s, k) = 0 \quad (3)$$

$$\frac{d}{dm} f(r, m, s, k) < 0 \quad (4)$$

$$\frac{d}{ds} f(r, m, s, k) > 0 \quad (5)$$

$$\frac{d}{ds} U(r, m, s, k) > 0 \quad (6)$$

[42] Continuity is considered for the standard metric. For $j, k \in \mathbb{N} \cap [1, N]$, hold all amounts tipped except t_j constant. Then as $t_j \to t_*$, $U\left(\left(t_1, \dots, t_j, \dots, t_N\right), t_k\right) \to U\left(\left(t_1, \dots, t_{j-1}, t_*, t_{j+1}, \dots, t_N\right), t_k\right)$.

[43] Note that t is the amount tipped by the consumer for whom this utility function applies. Of course, there is some value of n such that $t = t_n$, but including this level of specificity would be confusing.

$$\frac{d}{dk}U(r, m, s, k) > 0 \tag{7}$$

The first argument is supposed to represent the relationship between the server and the customer. In particular:

$$r = |C_1(\bar{t}, s) + C_2(\bar{t}, s)x - t| \tag{8}$$

In the above, I let x represent the size of the bill, suggesting that a linear relationship between the amount to be tipped and the size of the tipped left would result if people were only concerned with how much of a tip the server deserved.[44] However, I usually assume that everybody has a bill of the same size for the purpose of simplicity.

The second argument represents the desire to not spend money:

$$m = x + t \tag{9}$$

The third argument represents the desire to obtain as good service as possible. In particular, as by mechanisms previously discussed, I presume that customers attempt to influence service quality (if only subconsciously) through committing themselves in a somewhat visible manner to tip in a certain manner, hoping that doing so will incentivize the server to provide better service. Of course, this signaling process is not perfect, and should depend on the distribution of tips in the economy as well, as that distribution will affect the expectations of the servers. While I do not assign a specific form to the function that describes the third argument, I assume that whatever form that it may take should correspond with the following properties:

$$s(\bar{t}, t) \tag{10}$$

$$\frac{ds}{dt} > 0 \qquad \forall t \in [0, \infty) \tag{11}$$

[44] Alternatively, one could suggest that the function that determines how much one ought to tip is merely a smooth function, thus allowing for a linear approximation by using a cut-off Fourier series expansion, or even just a differentiable function that looks linear in a small enough domain. Practically speaking, the exact form of equation (6) is not important; what matters is that *r* increases in size as *t* gets farther away from an ideal value.

In other words, service is strictly increasing with the tip amount. I also want to say that service quality increases as the amount that anybody tips increases, as the server should perceive the incentive to provide good service as greater as the amount that people tip goes up.

$$\frac{\partial s}{\partial t_k} > 0 \qquad \forall k \in \mathbb{N} \cap [1, n], \forall t_k \in [0, \infty) \qquad (12)$$

In addition, it ought to be the case that service quality is always minimal when nobody tips at all:

$$s(0, t) = const. \qquad (13)$$

This last condition is meant to indicate the relative inability of a customer to improve his/her service through increased generosity of tipping when a tip is not expected.

Finally, the last argument in equation (1) represents the desire of the customer to be kind to the server. As such, this function should be monotonically increasing with respect to tip amount and service quality, except for when the tip amount is zero:

$$\frac{d}{dt} k(s, t, \bar{t}) \geq 0 \qquad \forall t \in (0, \infty) \qquad (14)$$

$$\frac{d}{ds} k(s, t, \bar{t}) \geq 0 \qquad \forall s \in [0, \infty) \qquad (15)$$

However, the kindness function should radically change character between a tip of zero and a positive amount. This is because a sufficiently small tip is perceived as insulting, whereas an amount to tip that is not insulting is determined by the average tipping behavior of the society in question. In fact, it is only in this regard that average societal behavior should affect the kindness function:

$$k(s, 0, \bar{t}) = \int_0^b \frac{d}{dt} k(s, t, \bar{t}) dt - k(s, b, \bar{t}) + \beta(\bar{t}) \qquad \forall b > 0 \qquad (16)$$

$$\frac{d}{d\bar{t}} k(s, t, \bar{t}) = 0 \qquad \forall \bar{t}, \forall t > 0 \qquad (17)$$

Furthermore, it should be the case that the more people tip, the less of an insult it should be to leave a tip. So say that β is a continuous function and that $\forall k \in \mathbb{N} \cap [1, n]$:

$$\frac{\partial}{\partial t^k} \beta(\bar{t}) \leq 0 \qquad \forall t^k > 0 \qquad (18)$$

Finally, I want to say that there is some amount that it would be absurd to tip more than. Specifically, let us say that there is some positive value T such that for any $\bar{t}, t > T$ implies that there is some $x \in [0, T]$ such that $U(\bar{t}, x) > U(\bar{t}, t)$. All this says is that nobody will ever find it to be in his/her interest to tip an infinite amount.

Let there be an infinite number of periods indexed by m, and say that s_m is a continuous function of $\bar{t}$ and t such that s_m uniformly converges to expression (10) over time (in the standard metric). This represents the process by which customers learn the effect of how much they tip on service quality. Let t_n^m be the amount that customer n tips in period m. Let U_m represent the utility function as the consumer understands it in period m. This can be expressed as:

$$U_m(\bar{t}, t) = U\big(r(\bar{t}, t), m(t), s_m(\bar{t}, t), k(s, t, \bar{t})\big)$$

In addition, $U_m \to U$ uniformly. The intuition behind this assumption is that one would hope that a small change in the service function would result in a small change in utility for any amount that one may tip. Furthermore, nobody tips an infinite amount, so we assume that for any $m \in \mathbb{N}$, the function U_m is maximized only on the interval $[0, T_m]$, where T_m is a positive number.

What results is the following theorem:[45]

Theorem 1: Given the setup in this section, assume that people tip an amount that maximizes U_m in each period m. If the social norm does not approach zero over time, then the amount tipped approaches a value at which utility is maximized. Formally, let t_m^n be the value that maximizes utility for consumer n in period m. Assume that everybody always tips a positive amount, and that U and all U_m are never maximized at zero. For any consumer n, for each convergent subsequence of the sequence $\{t_n^m\}$ in the standard metric space (and every strongly equivalent metric space), that subsequence converges to a point t_n^* in $[0, T]$. Furthermore, t_n^* maximizes $U(\bar{t}, t)$ given all other amounts tipped.

[45] Note that I am starting to abuse notation by writing my functions as functions of the amount tipped, rather than of these intermediate functions.

Thus, if there is one unique distribution of tips $\bar{t}$ at which $U(\bar{t}, t)$ is maximized for each consumer given the amount tipped by every other consumer, then $\bar{t}^m$ will converge to $\bar{t}$.

To prove this theorem, I need the following lemma:

Lemma: Let f_m be a sequence of continuous functions that converges uniformly to the continuous function f, each with the same domain of a closed interval. Let $\{x_m\}$ be a sequence of points such that for each $m \in \mathbb{N}$, x_m is a member of the argument that maximizes f_m. If the sequence converges, then it converges to a point in the argument that maximizes f.

Proof: First, note that the arguments in question always are nonempty because each of the functions in question is a continuous mapping from a closed domain. Let $\epsilon > 0$. Let x be the point to which the sequence $\{x_m\}$ converges (which is within the domain of these functions because the domain is closed). There is some δ such that whenever $|x - x_m| < \delta$, $|f(x) - f(x_m)| < \frac{\epsilon}{2}$. Furthermore, there is some value M_1 such that whenever $m > M_1$, $|x - x_m| < \delta$.

In addition, there is some value M_2 such that whenever $m > M_2$, $|f(x_m) - f_m(x_m)| < \frac{\epsilon}{2}$. Thus, by the triangle inequality, for any $m > \max\{M_1, M_2\}$:

$$|f(x) - f_m(x_m)| \leq |f(x) - f(x_m)| + |f(x_m) - f_m(x_m)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

This proves that $f_m(x_m) \to f(x)$. Now choose $\epsilon' > 0$. There is some number M_3 such that for any $m > M_3$, the maximum of f_m is within $\frac{\epsilon'}{2}$ of the maximum of f. Similarly, there is some number M_4 such that for any $m > M_4$, $|f(x) - f_m(x_m)| < \frac{\epsilon'}{2}$. Thus, by a similar application of the triangle equality, letting $m > \max\{M_3, M_4\}$, we can let F represent the maximum of the function f and say that:

$$|f(x) - F| \leq |f(x) - f_m(x_m)| + |f_m(x_m) - F| < \frac{\epsilon'}{2} + \frac{\epsilon'}{2} = \epsilon'$$

Since ϵ' is an arbitrary positive number, it follows that $f(x)$ is arbitrarily close to the maximum of f, and thus is the maximum of f. Therefore, x is a member of the argument that maximizes f. ■

Proof of Theorem: Fix n. All functions refer to consumer n as the consumer in question. This proof has five parts. The first part shows that the functions U and all U_m can be seen as functions only of t_n, and that in this regard $U_m \to U$ uniformly on any closed interval. The second part shows that, for any arbitrary $\delta > 0$, any convergent subsequence of $\{t_n^m\}$ has a subsequence in $[0, T+\delta]$. The third part applies the lemma. The fourth part shows that the sequence $\{t_n^m\}$ must have a convergent subsequence. The final part derives the final result.

First, note that any convergent subsequence of $\{t_n^m\}$ is part of a sequence in $\mathbb{R}^N$ of the amounts that each person tips in each of the relevant m. But this sequence in $\mathbb{R}^N$ also has a convergent subsequence. The sequence of nth components of this sequence is thus a convergent subsequence of the original subsequence in question. But for this subsequence of the convergent subsequence in question, all of the amounts tipped converge. For the rest of this proof, I will abuse notation by writing U and all of the U_m as functions only of t_n, letting all of the other t_j^m be given, as well as the t_j to which the t_j^m converge. But even as functions of only the activity of consumer n, note that that $U_m \to U$ uniformly.

Second, let $\delta > 0$. $\forall m \in \mathbb{N}$, U_m has a maximum on $[0, T+\delta]$. Let ϵ be the difference between the maximum of U and the supremum of the mapping of U from the set $(T+\delta, \infty)$. In the manner used twice in the proof of the lemma, for sufficiently large m, the maximum of U on $[0, T+\delta]$ and the maximum of U_m on the same interval must be within $\frac{\epsilon}{2}$ of each other, and the supremum of U and the supremum of U_m on the set $(T+\delta, \infty)$ must be within $\frac{\epsilon}{2}$ of each other. Thus, by the triangle inequality:

$$\max_{[0,T+\delta]} U_m - \sup_{(T+\delta,\infty)} U_m = \epsilon - \left[\max_{[0,T+\delta]} U - \max_{[0,T+\delta]} U_m\right] - \left[\sup_{(T+\delta,\infty)} U_m - \sup_{(T+\delta,\infty)} U\right]$$

$$\max_{[0,T+\delta]} U_m - \sup_{(T+\delta,\infty)} U_m \geq \epsilon - \left|\max_{[0,T+\delta]} U - \max_{[0,T+\delta]} U_m\right| - \left|\sup_{(T+\delta,\infty)} U_m - \sup_{(T+\delta,\infty)} U\right|$$

$$\max_{[0,T+\delta]} U_m - \sup_{(T+\delta,\infty)} U_m > \epsilon - \frac{\epsilon}{2} - \frac{\epsilon}{2} = 0$$

So for any arbitrary $\delta > 0$, for any large enough m, the argument that maximizes U_m is completely contained in $[0, T + \delta]$. Choose such a δ; for the convergent subsequence of $\{t_n^m\}$ in question, let M be sufficiently large such that all members of the subsequence with index larger than M are contained within $[0, T + \delta]$. Then all such points within the subsequence form a subsequence of the subsequence contained completely within $[-\delta, T + \delta]$.[46]

Third, I want to apply the lemma. Now note that on the domain $[0, T + \delta]$ the subsequence within this domain previously mentioned converges to a point, t_n^*, that is contained within the argument that maximizes U on the domain $[0, T + \delta]$, and thus on all of $\mathbb{R}$. Note that this subsequence in question is a subsequence of a convergent subsequence of $\{t_n^m\}$, and thus implies that the convergent subsequence of $\{t_n^m\}$ also converges to t_n^*, which maximizes U. In addition, note that $t_n^* \in [0, T]$ by the assumption expressed after equation (18).

Note that these results also apply for any metric strongly equivalent to the standard metric, as all such metrics would preserve the same properties of continuity and uniform continuity as exist in the standard metric.

Fourth, note that the sequence of amounts that all consumers tip, a vector in $\mathbb{R}^N$, must have positive components; in addition, I have already shown that for $\delta > 0$, for any consumer in question, for sufficiently large m, the argument that maximizes U_m is contained completely within $[0, T + \delta]$. Thus, any sequence of components has either T_m for some (finite) value m or T (since δ is an arbitrarily small number) as an upper bound. Thus, each sequence of componenets is bounded, and thus the entire sequence in $\mathbb{R}^N$ is bounded. So I can apply the Bolzano-Weierstrass Theorem to say that this sequence has a convergent subsequence. This subsequence converges in the manner described earlier in the proof.

[46] To keep track, in this proof we originally select any convergent subsequence. In the first part of the proof, we select a subsequence of this subsequence. Now, we are selecting a further subsequence.

Finally, suppose that there is only one vector of tip amounts in which the utility function in equilibrium is maximized for every consumer given the behavior of every other consumer. Specifically, consumer n maximizes U given that every other consumer is tipping the corresponding amount in the distribution when consumer n tips the nth component of the vector, t_n^*. From the rest of this proof, I know that t_n^* is the only accumulation point of the sequence $\{t_n^m\}$. The rest of the points must be isolated points. Let $\gamma > 0$, and note that the points in the convergent subsequence of $\{t_n^m\}$ that are within γ of t_n^* make up a subsequence that converges to t_n^*; the rest of the points in the sequence $\{t_n^m\}$ are isolated points; the set of these points cannot have t_n^* as an accumulation point, and thus have no accumulation point. But this set is bounded (because it is a subset of $\{t_n^m\}$, a bounded sequence), so by the contrapositive of the Bolzano-Weierstrass Theorem, this set must not be a sequence, and thus must have only a finite number of members. Let I be the largest index of the members of this set. Then every member of index greater than I is a member of the subsequence that converges to t_n^*; therefore, for any $\epsilon > 0$ there is a number $M > I$ such that $m > M$ implies that $|t_n^* - t_n^m| < \epsilon$. In conclusion, the sequence $\{t_n^m\}$ converges to the point t_n^*. ■

Define the social norm in period m as the average amount tipped in period m, i.e. $q_m = \frac{\sum_n t_n^m}{N}$.

Corollary 1.1: Under the conditions of Theorem 1, if every consumer has the same utility function, and if there is one vector of tips that could result from each consumer maximizing his/her utility given the amount each other consumer tips (one equilibrium point), then the equilibrium point is a vector in which each component is the same value, and over time the social norm approaches this value.

Proof: That the equilibrium point must be a constant times the identity vector is clear from the symmetry of every consumer having the same utility function; if distinct components j and k were to be swapped ($j, k \in \mathbb{N} \cap [1, N]$), this new vector would also be an equilibrium point, so the uniqueness of the equilibrium implies that these two vectors are the same, meaning that the two components must be equal. Thus, all components must be the same value, t^*.

The above theorem proves that, for any particular consumer, the sequence of values that the consumer tips will approach the value that makes up any arbitrary component of the equilibrium point; since the social norm is a continuous function of the tip amounts, it follows that the social norm approaches the average value of the components of the equilibrium point, which is t^*. ■

All of this demonstrates the rather intuitive result that, as a population of consumers with a sufficiently effective learning process in the manner described above learns, its behavior will increasingly approximate the behavior of the population in Nash equilibrium if the people initially had a perfect understanding. Of course, one could instead imagine that consumers cannot effectively keep up with the behavior of other consumers in real time. There are a couple of ways to model this phenomenon. One would be to presume that consumers observe the lagged behavior of their compatriots, and maximize utility assuming the amounts that each consumer tipped in the current period is the same as the amount tipped some number of periods previously. Yet the same proof as before would show that theorem 1 and its corollary would still hold.

The other method is the focus of this paper. In contrast to the learning process described above, in reality consumers likely do not perceive the changing patterns of tipping in the society around them, or at the minimum do not notice until looking back in hindsight. In the previous section, the representative consumer model led to the conclusion that ignorance of the effect that the improved average tip amount had on service quality led to a higher average tip amount, and thus a stronger norm for tipping. However, that result need not hold in this case.

Consider an example in which there are two customers in the economy. Imagine that one of the customers initially increases the amount that s/he (intends to) tip, while the other reduces that his/her corresponding (intended) tip amount. Furthermore, say that the wait staff is rational, but risk-adverse. Then the service quality would be a bit higher than the one who tipped less would expect when s/he does not take into account the amount tipped by the other customer, but the service quality is

significantly lower than the one who tipped more would expect by the corresponding oversight. The customer who previously tipped more may now tip less, while the customer who previously tipped less may now tip more. As the two customers continue to alternate, the one who tips less receives a minor lesson in how effective his/her (intended) tip amount is on service quality, while the other customer receives a harsh lesson on how ineffective his/her (intended) tip amount is on service quality. Finally, assume that the fluctuations in the amounts that the two customers tip diminish over time, with the understanding of the effects of their (intended) tip on service quality converging for the two customers. Then entry A5 of the appendix demonstrates that the amount that the two customers tip will eventually dip and remain below where it would have been if the two customers had perfect information regarding the effect of their (intended) tips on service quality, since the tip amounts will approach the Nash equilibrium in which both customers think that their tips have less of an effect on service quality than they actually have.

However, the result from the previous section *can* be recreated without the representative consumer model if one imposes some weaker assumptions. Suppose that, no matter what original understanding of the service function, s, customers may have, that they will eventually imagine that their tips have a greater effect on service than they actually do if they vary the amount that they tip and service quality *always* improves more with improved tips and worsens more with worse tips than would be suggested by the altered tip of the customer holding the rest of the tip amounts constant. Those that do not vary the amount that they tip must not be tipping at all, since they would have to vary the amount that they tip by continuity otherwise.[47] Then the following theorem holds:

Theorem 2: Under the above conditions, everybody who varies their tip amounts will tip more than they would if they knew the service function, s, if nobody tips less than the previous period when others tip

[47] Remember that utility is a continuous function of the service function when the tip amount is nonzero. Therefore, away from the zero tip amount, any change in the service function should lead to a different amount tipped when all other inputs are held constant.

more. At least as many people would not tip if they knew the service function, s, than if they did not know the service function.

Proof: If nobody tips less in the previous period when other people tip more, then the service quality will be improved by an amount greater than suggested by any one person's improved tip amount if the others are held constant in the case that some people improve their tip amounts. Similarly, if some people lessen the amount that they tip, nobody will increase the amount that they tip, and anybody who lessens the amount that they tip will find service quality to have plummeted more than it would have if only that one customer had reduced the amount that s/he tipped. Eventually, those that vary their tip amounts will learn that the amount that they tip will have a greater effect on service quality than it actually does.

Meanwhile, those who will not tip even upon thinking that their tips are more effective than they are would obviously not tip if they knew how ineffective their tips would be at generating improved service. Therefore, no fewer people tip under the learning condition than under the condition of perfect information. Anybody who tips under the learning condition who did not tip under the non-learning condition merely improves the service quality by either not approaching zero in tip amount or leaving zero in tip amount.

Thus, everybody who does not always refuse to tip thinks that their own tip amounts have a greater effect on service quality than they actually do. This implies that they will find it in their own interest to tip more, given the tips of all other customers, than they would if they had perfect information. The corollary of entry A5 of the appendix then implies that the tip amounts of those who tip positive amounts will approach a higher value than they would tip in equilibrium under perfect information.

Finally, for some given customer, let t^L represent the amount that s/he tips under the learning condition. Let t^I be the amount that the same customer would tip under the perfect information

condition. Note that $t^L > t^I$. For any period $n \in \mathbb{N}$, let t_n^L represent the amount that the customer tips in period n under the learning condition, and let t_n^I represent the amount that the customer tips in period n. Since $t_n^L \Rightarrow t^L$ and $t_n^I \Rightarrow t^I$, we can let $\epsilon = t^L - t^I > 0$ and note that there is a natural number $N \in \mathbb{N}$ such that $n > N$ implies that $t^L - t_n^L \leq \frac{|t^L - t^n|}{2} < \frac{\epsilon}{2} = \frac{t^L - t^I}{2} \Rightarrow t_n^L > \frac{t^L + t^I}{2}$, and also that $t_n^I - t^I \leq |t_n^I - t^I| < \frac{\epsilon}{2} = \frac{t^L - t^I}{2} \Rightarrow t_n^I < \frac{t^L + t^I}{2}$. Thus, for $n > N$, $t_n^L > t_n^I$. ■

If one allowed for some to increase the amount that they tip while others decrease the amount that they tip, then one would need to further specify the learning process in order to see what would happen to the average tip amount over time. Intuitively it makes sense that the average tip would still be greater under the learning condition than under perfect information as long as there are usually an overwhelming majority changing the amount that they tip in the same direction. However, one could imagine a situation in which servers are particularly risk-averse, and act as if only the lowest amount tipped is the tip they could receive for good service. In that case, there is plenty of reason to suspect that the average amount tipped would plummet under the learning condition as long as the smallest amount tipped never increased.

Finally, note that the sticky zero principle was demonstrated in the last section. The example used in the previous section still applies, since this section is a generalization of the representative consumer model used in section IV.

Section VI: The Societal Welfare Effects of the Norm of Tipping

A multitude of authors have argued that the social norm of tipping is ultimately bad for societal welfare, as customers often feel coerced to leave a tip against their wishes, and similarly that the norm of tipping largely arises from a ruinous competition between patrons for status. In addition, some claim that tipping results in the proliferation of prejudice and facilitates tax evasion. While these are legitimate critiques, the argument against this widespread custom has significant weaknesses. Even given the inherent flimsiness of discussing welfare effects of behavior that is clearly largely motivated by nonmaterial conditions, it seems clear that tipping incentivizes good service to a sufficient degree that employers see it as an effective means by which to motivate their servers. Under reasonable labor market conditions, the welfare of servers should not depend on the norm of tipping. Presuming that the expectation of tipping does not cause immense suffering among customers, the welfare effects of the tipping norm are ambiguous at best.

Employers voluntarily choose to employ the tipping system, since they could otherwise impose a service charge and thus discourage customers from tipping the wait staff. There is a general misperception that one should not tip when there is a service charge, as indicated in Schein, Jablonski, and Wohlfahrt (1984, p. 14). This feeling largely arises from the sentiment that the service charge already goes to the servers, which need not be the case.

Therefore, if the norm of tipping is bad for anybody, presumably it is bad for the customers. Margalioth (2006) argues that tipping is bad for society because of tax evasion and the fact that better tips from wealthier customers may put pressure on those customers less well-off to tip better, an externality imposed on the customers. The first reason is not an inherent problem with the practice of tipping, but with the way that the government collects tax revenue. The second reason is an argument against consumption similar to that made in Frank (2001). It could be argued that the externality that the upper class imposes on the middle class is bad for societal welfare.

However, tipping is usually based on the size of the bill, and people usually go to restaurants based on their own personal wealth. Furthermore, Frank (2001) argues that spending by the wealthy in general puts pressure on the less wealthy to spend more ("keeping up with the Joneses"). It is by no means clear why tipping should be signaled out for this phenomenon.

Azar (2005) argues that tipping is an innovation in order to escape the trust game, and as such is beneficial for societal welfare. The BDPD model agrees with his analysis. The reason why the norm developed was to incentivize servants in Tudor England; people still tip more when service is better and servers know it, creating an informal market for service quality where none formally exists.

Of course, this is not meant to disregard these complaints regarding the societal welfare implications of the social norm of tipping. After all, it could be the case that, in contrast to this model's exposition, people merely feel pressured in a very disconcerting way to tip when they do not really want to. But this also begs the question of why these customers wouldn't simply avoid those restaurants that employ a policy of encouraging tipping. Furthermore, restaurants in a competitive market should seek to employ a service charge if tipping reduced welfare. The challenge is now for those who think that the norm of tipping is bad for societal welfare to justify this claim in a significantly more rigorous manner.

However, the BDPD model does not rule out the possibility that such a justification could be found. Social norms do not just arise from resolutions to trust games and prisoner's dilemmas; they also come from those in positions of power imposing values that benefit their interests, which may or may not also be the general interests of those in society. It could be that tipping is a particularly insidious way that the wealthy compel others to spend money that they otherwise wouldn't. Perhaps the wealthy also accrue status from others by tipping more than them, shaming others more than they benefit from the increased status. A model could be constructed in which that turns out to be the case; however, such a model would have to show what makes the pressure to tip more different from the generic pressure to spend more.

Section VII: Implications of the Biologically-Determined Planner-Doer Model

The social norm of tipping can be completely understood form the perspective of the biologically-determined planner-doer (BDPD) model. However, there is no reason why the BDPD model could not be applied to other significant social norms, even though the focus may extend beyond subjects traditionally considered under the domain of economics. The BDPD model has the potential to thoroughly explain such phenomena as the origin of social constructs such as race and gender, the rise of belief systems, and the promulgation of hierarchies.

Section II hinted that the BDPD model could explain why cooperation sometimes occurs in the prisoner's dilemma. Appendix entry A6 goes into detail as to how the BDPD model predicts that all combinations of cooperation and betrayal are possible in the iterated prisoner's dilemma. For the single-round prisoner's dilemma, the BDPD model works in a straight-forward manner. Let each player be able to send a signal to the other player. For each player, there is a probability p that a player who sends that signal and receives the signal in return will be pre-committed to choosing cooperation in the prisoner's dilemma. Each player already knows whether or not s/he can send a false signal. Assume that a player will send a false signal whenever possible. If p times the benefit from having the other player cooperate is greater than the cost of cooperating, then both players will send the signal whether or not either of them can falsely signal. If the probability is too low, then both players will automatically choose betrayal.

Obviously, it is better for societal welfare if the signal is stronger, i.e., if p is closer to one. In fact, everybody in a society has a personal interest in weeding out those who are more capable of sending a false signal in order to increase the strength of the signal. To see this, note that people only choose to cooperate because they send the signal, and they do not bother sending the signal unless they think that there is a good enough chance of solving the prisoner's dilemma in order to warrant the risk.

Since they could always choose betrayal by not signaling, it is clearly in each player's interest to see to it that the signal is as strong as possible.

Yet there is a limit to the desire that everybody in the society would have to increase the strength of the signal. If it is impossible for the signal to become strong enough for cooperative equilibria to arise, it will not be worthwhile to spend effort eliminating those who are more capable of sending false signals. If it takes any effort at all to maintain a positive value for p, then the lack of effort will send p to zero, implying a complete lack of trust in the society. Thus, using a fixed strength for the signal, it appears that there is a level of civic trust which must be attainable, and thus attained; otherwise, that trust goes away completely.[48]

Those in positions of power are more capable of organizing society in such a way as to weed out those who can falsely signal; this motivates those without power to support the creation of social hierarchies. Those in power have a powerful interest in the welfare of society in general because they generally exact rents from society through the exercise of their power. Therefore, those in power will tend to propagate societal values that are beneficial for society as a whole. Of course, they can also propagate values that justify and defend the hierarchy that provide them their power. In this sense, the social hierarchy is created by a social contract in that most people benefit from the establishment of that hierarchy.

Since specialization tends to be beneficial for society as a whole, the empowered members of society will tend to support specialization through the system of values that they impose. But the empowered tend to play different roles than others do. Thus, the empowered have an interest in glorifying their roles in order to enhance their status and authority, simultaneously reducing the status and authority of everyone else. Social constructs can easily arise from these different roles. For

[48] It could readily be argued that one would know whether or not others could send false signals based on past experience. However, in a large enough society, it is unlikely that one would be playing the prisoner's dilemma with the same person multiple times.

example, gender roles may be economically efficient in certain pre-modern economies, but the men are usually empowered, and thus have an interest in elevating the status of men and degrading the status of women.

However, other times social constructs can serve the role of getting those on the lower end of the hierarchy to become further invested in maintaining the hierarchy. Racism in the United States could pit lower-class white people against black people, keeping wealthy white Americans in a position of power by keeping everyone else divided. The ultra-low status classes of pre-modern India and Japan give the peasants a scapegoat and a group of people that they can feel better than, thus keeping them more content with the status quo.

One particularly interesting application of the model is in the process of voting in an election with a large population. If one imagine an election in which one hundred million people vote between two candidates, each person is equally likely to vote for one candidate or the other, and the probability of one person voting for one candidate is independent of the votes of any other voter, then the probability that one person's vote will make the difference is .000079788.[49] If one made eight dollars an hour and it took one an hour to vote, then the discounted utility of choosing the winner of that election ought to be greater than the utility gained from an immediate payment of one hundred thousand dollars in order to make it worthwhile to vote. Considering the fact that eight dollars is barely above the minimum wage, the average income in the United States is well less than one hundred thousand dollars a year, and over a hundred million people vote for president, it appears that over a hundred million people in the United States are voting irrationally. This has motivated Nelson and Greene (2003) to use a signaling process similar to the BDPD model to explain voting.

Finally, there is no particular reason why social norms could not include changing one's belief system. It may be that some in society would benefit significantly from believing in a certain way, or at

[49] The program used to arrive at this calculation in in appendix A8.

least convincing others that they believed that way. A rational actor could just pretend to believe while it was in his/her best interests. But if the other members of society are looking for a signal that is hard to fake, belief systems can be propagated throughout most of the population because those beliefs benefit everybody in the society; anybody refusing to believe could be inflicting a negative externality upon everybody else. A clear example of this was shown in section IV, in that the amount tipped was greater when people thought their tips had a greater impact on service quality than they actually did. Note that service quality improves for everybody when one person tips more; therefore, everybody tipping as if their tips alone were responsible for the entire improvement in service is a far more efficient result for society than one in which people know better and tip less. The increased tips compensate for the positive externality that tipping has for the service quality for all customers. Therefore, society as a whole has a powerful interest in promoting everybody to think that their tips are far more effective at improving service quality than they actually are.

Section VIII: Conclusion

The biologically-determined planner-doer model does an excellent job in explaining the tipping norm, and furthermore suggests a myriad of explanations for a variety of phenomena whose origins are puzzling from the perspective of the rational actor. Yet the complexities of possibilities that result from the model suggest that intellectual laziness must be avoided at all costs. History matters when it comes to the development of social norms, and there is no clear-cut way to tell when there are and are not social norms that come into play when examining economic questions.

The BDPD model doesn't explain everything. Since the doer still maximizes utility, deviations from rationality such as the Allais paradox still pose a problem, although evolutionary consideration may offer a way out. Furthermore, the model still relies on a signaling process that needs to be better understood before it can be modeled in a more sophisticated manner. Yet the model holds the promise of understanding a much greater variety of economic phenomena than would be possible with the model of the rational actor.

Bibliography

Akerlof, George A., and William T. Dickens. "The Economic Consequences of Cognitive Dissonance." *The American Economic Review* 72.3 (1982): 307-319.

Arrow, Kenneth. "Political and Economic Evaluation of Social Effects and Externalities," in M. Intriligator, ed., *Frontiers of Quantitative Economics* (Amsterdam: North-Holland, 1971), 5-25.

Ariely, Dan. *Predictably Irrational, Revised and Expanded Edition: The Hidden Forces that Shape our Decisions.* Harper, 2009.

Ayres, Ian, Fredrick E. Vars, and Nasser Zakariya. "To Insure Prejudice: Racial Disparities in Taxicab Tipping" in *The Yale Law Journal*, Vol. 114, No. 7 (May 2005), pp. 1613-1674. Published by: The Yale Law Journal Company, Inc. http://www.jstor.org/stable/4135761

Azar, Ofer H. "What Sustains Social Norms and How They Evolve? The Case for Tipping." Journal of Economic Behavior and Organization 54 (May 2004), 49–64. http://www.sciencedirect.com/science/article/pii/S016726810300221X#

Azar, Ofer H. "Optimal Monitoring with External Incentives: The Case for Tipping" in *Southern Economic Journal*, Vol. 71, No. 1 (July 2004), pp. 170-181. http://www.jstor.org/stable/4135319

Azar, Ofer H. "The Social Norm of Tipping: Does it Improve Social Welfare?" Journal of Economics 85.2 (2005): 141-173. http://link.springer.com/article/10.1007/s00712-005-0123-0#page-1

Azar, Ofer H. "Relative Thinking Theory" in *Journal of Socio-Economics*, Volume 36, Issue 1, February 2007, Pages 1–14. http://dx.doi.org/10.1016/j.socec.2005.12.014

Azar, Ofer H. "Do People Tip Because of Psychological or Strategic Motivations? An Empirical Analysis of Restaurant Tipping." *Applied Economics* 42.23 (2010): 3039-3044. https://trunk.tufts.edu/access/content/attachment/5ea012a6-f228-4a85-991aa0715436a24a/Syllabus/a2d15526-61dc-40f3-aeb5-c4d86225366f/azar.pdf

Azar, Ofer H. and Yossi Tobol. "Tipping as a Strategic Investment in Service Quality: An Optimal-Control Analysis of Repeated Interactions in the Service Industry.: *Southern Economic Journal*, Vol. 75, No. 1 (July, 2008), 246-260. Published by: Southern Economic Association. Stable URL: http://www.jstor.org/stable/20112038

Box, George EP, and Norman R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley. High Occupancy Vehicle 4 (1987).

Carrión, Ricardo E., Julian P. Keenan, and Natalie Sebanz. "A Truth that's Told with Bad Intent: an ERP Study of Deception." *Cognition* 114.1 (2010): 105-110. http://somby.info/page3/assets/2010_percent20Carrion.pdf

Cesarini, David, Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. "Heritability of Cooperative Behavior in the Trust Game." Proceedings of the National Academy of sciences 105.10 (2008): 3721-3726. http://jhfowler.ucsd.edu/heritability_of_cooperative_behavior.pdf

Cheyney, Edward P. "The Disappearance of English Serfdom." The English Historical Review 15.57 (1900): 20-37. http://www.jstor.org/stable/pdfplus/548409.pdf?acceptTC=true

Cho, Minho. "A Re-Examination of Cultural Influences on Restaurant Tipping Behavior: A Comparison of Japan and the US." Journal of Foodservice Business Research 8.1 (2006): 79-96. http://www.tandfonline.com/doi/pdf/10.1300/J369v08n01_06

Elster, Jon. "Social Norms and Economic Theory" in *The Journal of Economic Perspectives* , Vol. 3, No. 4 (Autumn, 1989), pp. 99-117Published by: American Economic Association Article Stable URL: http://www.jstor.org/stable/1942912

Frank, Robert H. "If Homo Economicus Could Choose his own Utility Function Would he Want One with a Conscience?" The American Economic Review (1987): 593-604.

Frank, Robert. *Passion Within Reason: The Strategic Role of the Emotions*. New York: Norton (1988).

Frank, Robert H. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. Free Press (2001).

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. "Does Studying Economics Inhibit Cooperation?." The Journal of Economic Perspectives (1993): 159-171. http://psych.cornell.edu/sec/pubPeople/tdg1/Frank,Gilo,Regan.93.pdf

Fehr, Ernst, and Klaus M. Schmidt. (2001): Theories of Fairness and Reciprocity. *Evidence and Economic Applications.* Discussion Papers in Economics 2001-2.http://epub.ub.uni-muenchen.de/14/

Falk, Armin, and Ernst Fehr. "Psychological Foundations of Incentives" in *European Economic Review*, Vol. 46, No. 4-5 (May, 2002), pp. 687-742. http://dx.doi.org/10.1016/S0014-2921(01)00208-2

Glimcher, P. W., Dorris, M. C. & Bayer, H. M. "Physiological Utility Theory and the Neuroeconomics of Choice." Games Econ. Behav. 52, 213–-256 (2005).

Greene, Kenneth V., and Phillip J. Nelson. *Signaling Goodness: Social Rules and Public Choice*. Ann Arbor: The University of Michigan Press (2003).

Greenfeld, Liah. The Spirit of Capitalism: Nationalism and Economic Growth. Harvard University Press, 2001.

Gregg, Aiden P. "When Vying Reveals Lying: The Timed Antagonistic Response Alethiometer." Applied Cognitive Psychology 21.5 (2007): 621-647. http://onlinelibrary.wiley.com/doi/10.1002/acp.1298/pdf

Hane, Mikiso. *Premodern Japan: a Historical Survey*. Boulder, Colorado: Westview Press, 1991.

Harris, Judy, Michael Lynn and George M. Zinkhan. "Consumer Tipping: A Cross-Country Study" in *Journal of Consumer Research*, Vol. 20, No. 3 (Dec., 1993), pp. 478-488. http://www.jstor.org/stable/2489361

Hodgkin, Alan L., and Andrew F. Huxley. "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve." The Journal of Physiology 117.4 (1952): 500. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413/pdf/jphysiol01442-0106.pdf

Jansen, Arthur SP, et al. "Central Command Neurons of the Sympathetic Nervous System: Basis of the Fight-or-Flight Response." Science 270.5236 (1995): 644-646. http://www.sciencemag.org/content/270/5236/644.full.pdf

Lange, Oskar. "On the Economic Theory of Socialism: Part Two." *The Review of Economic Studies*, Vol. 4, No. 2 (Feb., 1937), pp. 123-142. http://www.jstor.org/stable/pdfplus/2967609.pdf

Levitt, Steven D., John A. List, and David H. Reiley. "What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments." Econometrica 78.4 (2010): 1413-1434. http://www.fieldexperiments.com/papers/00080.pdf

List, John A. "Does Market Experience Eliminate Market Anomalies?" The Quarterly Journal of Economics 118.1 (2003): 41-71. http://karlan.yale.edu/fieldexperiments/papers/00297.pdf

List, John A. "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace." Econometrica 72.2 (2004): 615-625. http://karlan.yale.edu/fieldexperiments/papers/00174.pdf

Lynn, Michael. "Black-White Differences in Tipping of Various Service Providers." Journal of Applied Social Psychology 34.11 (2004): 2261-2271. http://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2004.tb01976.x/pdf

Lynn, Michael, and Benjamin Katz. "Are Christian/Religious People Poor Tippers?" Journal of Applied Social Psychology (forthcoming; 2011). http://tippingresearch.com/uploads/ChristianTippersJASPaccepted.pdf

Lynn, William. "Tipping in Restaurants and Around the Globe: An Interdisciplinary Review." (2006). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=465942

Margalioth, Yoram. " The Case against Tipping." U. Pa. J. Lab. & Emp. L. 9 (2006): 117. http://heinonline.org/HOL/Page?handle=hein.journals/upjlel9&div=10&g_sent=1&collection=journals

Nelson, Phillip J., and Kenneth V. Greene. *Signaling Goodness: Social Rules and Public Choice*. University of Michigan Press, 2003.

Ostrom, Elinor. "Collective Action and the Evolution of Social Norms." The Journal of Economic Perspectives 14.3 (2000): 137-158. http://www.jstor.org/stable/pdfplus/2646923.pdf?acceptTC=true

Padoa-Schioppa, C. & Assad, J.A. "Neurons in the Orbitofrontal Cortex Encode Economic Value." Nature 41, 223–226. http://www.nature.com/nature/journal/v441/n7090/full/nature04676.html

Parrett, Matt. "An Analysis of the Determinants of Tipping Behavior: A Laboratory Experiment and Evidence from Restaurant Tipping." *Journal of Socio-Economics* 29:203-14 (2006). Stable URL: http://www.jstor.org/stable/20111903

Rind, B., & Strohmetz, D. "Effect of Beliefs about Future Weather Conditions on Restaurant Tipping." Journal of Applied Social Psychology, 31, 2160–2164 (2001). http://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2001.tb00168.x/pdf

Roth, Alvin E., and J. Keith Murnighan. "Equilibrium Behavior and Repeated Play of the Prisoner's Dilemma." Journal of Mathematical psychology 17.2 (1978): 189-198. http://ac.els-cdn.com/0022249678900305/1-s2.0-0022249678900305-main.pdf?_tid=f6cc3228-adeb-11e2-8c25-00000aab0f26&acdnat=1366924154_e9c5c6da7d4c164dd01a275d97e9facc

Segal, Nancy L., and Scott L. Hershberger. "Cooperation and Competition Between Twins: Findings from a Prisoner's Dilemma Game." Evolution and Human Behavior 20.1 (1999): 29-51. http://www.sciencedirect.com/science/article/pii/S1090513898000397

Segrave, Kerry. *Tipping: An American Social History of Gratuities.* Jefferson: McFarland, 1998.

Schein, John E., Edwin F. Jablonski, and Barbara R. Wohlfahrt. *The Art of Tipping: Customs & Controversies.* Wausau, WI: Tippers International, 1984.

Sinn, Hans-Werner. "Weber's Law and the Biological Evolution of Risk Preferences: The Selective Dominance of the Logarithmic Utility Function, 2002 Geneva Risk Lecture." The Geneva Papers on Risk and Insurance Theory 28.2 (2003): 87-100. http://link.springer.com/article/10.1023/A percent3A1026384519480

Thaler, Richard H., and H. M. Shefrin. “An Economic Theory of Self-Control.” Journal of Political Economy,Vol. 89, No. 2 (Apr., 1981), pp. 392-406. http://www.jstor.org/stable/1833317

Appendix A1: Proofs Related to the Model in Section IV

For the following proofs, all variables independent of *t* are given, so we say that $U\big(r(t), m(t), s(t), k(t)\big) = U(t)$ and $U\big(r(t), m(t), s_n(t), k(t)\big) = U_n(t)$ for short.

Proof of Lemma 1:

Suppose all variables not dependent on *t* are given. Since the utility function is continuous, there is a value $t \in [0, T]$ for *T* as described in the lemma such that the utility function never is greater than $U(t)$ for any values within $[0, T]$. Denote this value by t^*. Let $t \in [T, \infty)$. By the mean value theorem, there is some value $c \in (T, t)$ such that:

$$U(T) - U(t) = U'(c)(T - t)$$

But note that equation (12) of the model implies that:

$$U'(c) \leq 0 \qquad \forall c \in [T, \infty)$$

Also note that $T - t \leq 0$. Thus:

$$U(t^*) \geq U(T) \geq U(t) \quad \forall t \in [T, \infty)$$

Since $t \in [0, \infty)$, it follows that t^* is the global maximum. ■

Proof of Theorem 1:

Let $\{\delta_m\}$ be a positive real-valued sequence such that $\{\delta_m\} \to 0$. Let $t^m = \{argmax\_\text{t U(t)} | t \notin (t^* - \delta_m, t^* + \delta_m)$. Note that whenever $U(x) > U(t^m)$, it must be the case that $x \in (t^* - \delta_m, t^* + \delta_m)$. Fix $m \in \mathbb{N}$. Let $\epsilon = U(t^*) - U(t^m) > 0$. Then $\exists N \in \mathbb{N}$ such that $\forall n > N$, $|U(x) - U_n(x)| < \frac{\epsilon}{2}$ $\forall x \in \mathbb{R}$. By the triangle inequality:

$$U(t^*) - U(t_n) = |U(t^*) - U(t_n)| \leq |U(t^*) - U_n(t^*)| + |U(t_n) - U_n(t_n)| < \epsilon = U(t^*) - U(t^m)$$

Thus, for fixed *m*, for sufficiently large *n*, $U(t_n) > U(t^m)$.

Let $\delta > 0$. Because $\delta_m \to 0$, $\exists M \in \mathbb{N}$ such that $\forall m > M$, $\delta_m < \delta$. But note that the previous result implies that t_n is within δ_m of t^* when *n* is sufficiently large, implying that t_n is within δ_m of t^* when *n* is sufficiently large. Therefore, $t_n \to t^*$. ■

Proof of Theorem 2:

The proof that $t_n \to t^I$ is identical to the one used in theorem 1. From here, note that $\bar{t} = t$ and:

$$\frac{dS}{dt} = \frac{\partial s}{\partial \bar{t}} + \frac{\partial s}{\partial t}$$

This and equations (7) and (14) from the model imply that $\forall t \in [0, \infty)$:[50]

$$\frac{dP}{dt} - \frac{dU}{dt} = \left(\frac{\partial P}{\partial r}\frac{\partial r}{\partial t} + \frac{\partial P}{\partial m}\frac{\partial m}{\partial t} + \frac{\partial P}{\partial s}\left(\frac{\partial s}{\partial \bar{t}} + \frac{\partial s}{\partial t}\right) + \frac{\partial P}{\partial k}\frac{\partial k}{\partial t}\right) - \left(\frac{\partial U}{\partial r}\frac{\partial r}{\partial t} + \frac{\partial U}{\partial m}\frac{\partial m}{\partial t} + \frac{\partial U}{\partial s}\frac{\partial s}{\partial t} + \frac{\partial U}{\partial k}\frac{\partial k}{\partial t}\right)$$

$$\frac{dP}{dt} - \frac{dU}{dt} = \left(\frac{\partial U}{\partial r}\frac{\partial r}{\partial t} + \frac{\partial U}{\partial m}\frac{\partial m}{\partial t} + 2\frac{\partial U}{\partial s}\frac{\partial s}{\partial t} + \frac{\partial U}{\partial k}\frac{\partial k}{\partial t}\right) - \left(\frac{\partial U}{\partial r}\frac{\partial r}{\partial t} + \frac{\partial U}{\partial m}\frac{\partial m}{\partial t} + \frac{\partial U}{\partial s}\frac{\partial s}{\partial t} + \frac{\partial U}{\partial k}\frac{\partial k}{\partial t}\right)$$

$$\frac{dP}{dt} - \frac{dU}{dt} = \frac{\partial U}{\partial s}\frac{\partial s}{\partial t} > 0$$

Let $t^c < t^*$. Then by the second part of the Fundamental Theorem of Calculus:

$$P(t^*) - P(t^c) = \int_{t^c}^{t^*} \frac{dP}{dt} dt > \int_{t^c}^{t^*} \frac{dU}{dt} dt = U(t^*) - U(t^c) > 0$$

$$P(t^*) > P(t^c)$$

Thus, $t^I \geq t^*$. Furthermore:

$$\frac{dP}{dt}(t^*) > \frac{dU}{dt}(t^*) = 0$$

This implies that *P* is not optimized at t^*, implying that $t^I \neq t^*$. Thus, $t^I > t^*$.

Finally, let $\epsilon = t^I - t^*$. Then $\exists N$ such that $\forall n > N$, $t^I - t_n < \epsilon = t^I - t^*$, implying that:

$$t_n > t^* \qquad \blacksquare$$

Proof of Theorem 3:

The proof for most of this theorem parallels the proofs for theorems 1 and 2. Simply note that the discontinuity has no effect when the average amount tipped is above zero, so the proofs for theorems 1 and 2 hold. Now, assume that r is always zero, while $m \equiv -k - t$ and $U = r + k + m + S_n$.

[50] Technically this ignores the fact that the relationship derivative should be different between the two functions. However, note that this should in fact reinforce the conclusion, as in the new function an increase in the tip increases the service quality faster, thus increasing the pressure to tip more.

Note that this simplifies to $U = S_n - t$. When the average amount tipped is zero, tipping will lead to no improvement in service quality, so the amount one should tip that maximizes utility is zero. Note that this continues in perpetuity, and furthermore that $S \equiv \lim_{n \Rightarrow \infty} S_n \equiv 0$. ■

Appendix A2: A Brief Justification for Learning through Uniform Convergence

When it came to describing the method by which customers learned how service depended on various conditions, I asserted that the function being learned should be modeled by a sequence of functions $\{f_n\}$ such that $\{f_n\} \to f$ uniformly. Admittedly, this need not be the case. Suppose there is some number $\epsilon > 0$ such that a creature will survive if within *N* periods of time, and for some $\vec{x}$ yet to be revealed, the creature could guess the value of $f(\vec{x})$ with ϵ accuracy. In other words, the creature will survive if and only if the following condition holds:

$$|f(\vec{x}) - f_N(\vec{x})| < \epsilon$$

Uniform convergence assures that there is such a value of *N* such that this condition holds, and one must simply hope that this value of *N* will do. In fact, let's suppose it does; furthermore, let's suppose for simplicity that the all of these functions map the domain $[0,1]$ to $\mathbb{R}$. So $\forall x \in [0,1]$, we have that:

$$\forall n \geq N, |f(x) - f_n(x)| < \epsilon$$

But suppose that a second creature learns through a sequence of functions $\{g_n\}$ such that $g_n(x) = f_n(x)$ whenever $x \notin \left\{x | \frac{1}{x} \in \mathbb{N}\right\}$. Furthermore, suppose that the probability that any value is chosen from the domain is the same as the probability of any other value being chosen. Then:

$$P(|f(x) - g_N(x)| < \epsilon) = 1$$

Thus, the second creature is just as likely to survive as the first, even though the second creature doesn't learn through the utilization of a sequence of uniformly convergent functions.

However, this argument is of no practical significance; one can merely note that learning should occur through a sequence of functions that uniformly converge on a set of values such that the probability that a given value in the domain is contained within the set in question is one. Then ignoring the set of other possibilities and claiming uniform convergence is simply a matter of ignoring a set of possibilities with probability zero, a quite reasonable simplification. For example, if one were to model the learning process of the second creature in the above example, it would be reasonable to claim that

the creature learned through a uniformly convergent sequence of functions, as the probability that a test value will occur on the set on which uniform convergence doesn't hold is zero.

Yet fundamentally, the main reason to assume that such a learning process is in play is that one ought to assume that the empirical process can reveal the fundamental nature of how observable phenomena occur, as this is the basis of scientific inquiry. If one supposes this to be the case, then one should be able to take a sufficiently accurate guess of a function that describes an observable phenomenon after enough observation. In other words, there should be a sequence of functions $\{f_n\}$ such that for any $\epsilon > 0$ there is an *N* such that for any $n > N$:

$$|f(x) - f_n(x)| < \epsilon \; \forall x$$

This is the definition of uniform convergence.

Appendix A3: Model Demonstrating Instability of Soft Social Norms over Time

Subsection I: Discussion of Model

The following is a model of tipping in which the social norm is determined as an average amount tipped in the previous period, and every individual feels pressured to conform to the social norm despite their personal preferences. Although this model is being developed in the context of tipping, it should apply to any soft social norm, i.e. any norm that is not perfectly clearly defined. For example, while the correct amount to tip can be murky at times, the social norm of monogamy is quite clear. One could make a logically consistent argument that even if nobody was personally in favor of purely monogamous marriages, nobody would be willing to suffer the social consequences of marrying even one extra person, thus maintaining the social norm of monogamy. However, it would not make sense to argue that if everyone wanted to tip nothing, that nobody would do things such as round down to the nearest dollar (or even ten cents), or judge service by a slightly harsher standard than had been done in the past. Once people behave in such a manner, the norm should decline over time. Thus, soft social norms such as tipping cannot be maintained purely through a desire to copy the behavior of others, although sharply defined norms such as monogamy could be.

The intuition behind the model described below is relatively simple. If the social norm is not too sharply defined, it should be the case that people seek to tip slightly below the average of the previous period in order to save money without looking like clear deviants from the norm. Thus, a tip may be calculated before tax rather than afterwards, decimal values rounded down rather than up, or simple guesses as to the correct amount to tip tending to lie on the lower side of acceptable answers. If there is any range of values that one could tip that could be considered acceptable, people should tend to pick the lower values, and thus the norm should decline over time.

What is far more surprising is the need for the norm to be "soft" enough so that people are able to tip slightly below the average of the previous period. A well-defined norm can create such a strong

incentive to tip at the average of the previous period that nobody would dare to defy the norm thus established. Obviously this is an unrealistic situation, as there will almost always be rebels who will resist social norms no matter how much pressure there is to conform to them. However, if such people can be effectively ostracized or otherwise indicated as beyond social approbation, which should be possible if the social norm is sharply defined such that it is clear when people are not conforming to it exactly, then it could well be the case that people are only exposed to or care about conforming to the behavior of those who conform to the norm. In such a situation, the relevant question to the survival of that norm becomes whether or not the population experiences positive population growth, including the ostracized renegades among those who have left the population in question. If a population does not expel or otherwise disregard those who violate the norm, then of course the norm will decline over time. However, given that there are always conformists in the population, the norm may only approach zero; it's possible that the norm may never reach zero, despite the inclusion of nonconformists.

The model is an improvement on the model developed in Azar (2004a), relaxing many assumptions from the previous model so that the model can be applied to a myriad number of situations in which social norms have economic impacts. However, this model inherits from its predecessor a serious flaw in that it presumes that people feel a pressure to conform exactly to the average behavior of the previous period. Instead, it could very well be the case that society creates a social norm to tip merely by encouraging a perception that tipping is inherently good, and the more that is tipped, the better. Thus, the implications of this model are limited; the model developed in the main body of this paper is far more useful for more general applications.

In the next subsection, I model the deterioration of soft social norms over time. In the subsequent subsection, I briefly examine the case of sharp social norms. Proofs follow in the appendix.

Subsection II: The Model in Detail

I construct an infinite-period model with in which everybody attempts to maximize their utility, which is a function of the tip that they give and the social norm for tipping. Everybody has the same utility function, and everybody chooses the same individual tip in any given period. The social norm for any period is determined by the average tip from the last period. Assume that the utility function is defined as the sum of two continuous functions, the individual function and the social function. From either side of any social norm for tipping, the value of the social function monotonically increases as the individual tip approaches the social norm for tipping. Similarly, there is some ideal tip such that from either side of the ideal tip, the individual function monotonically increases as the individual tip approaches the ideal tip. If we denote the ideal tip by t_*, the time period by k, the individual tip in that time period by t_k, the individual function by $I(t_k)$, the social norm for tipping in period k by n_k, the social function by $S(t_k,n_k)$, and the utility function by $U(t_k,n_k)$, then one can express these conditions mathematically as follows:

$$t_k = \text{argmax}_x\, U(x, n_k) \tag{1}$$[51]

$$t_k = n_{k+1} \tag{2}$$

$$U(t_k, n_k) = I(t_k) + S(t_k, n_k) \tag{3}$$

$$\forall a, b \in \mathbb{R}, |a - n_k| < |b - n_k|, |a - b| < \max\{|a - n_k|, |b - n_k|\} \Rightarrow S(b, n_k) \leq S(a, n_k) \tag{4}$$

$$\forall a, b \in \mathbb{R}, |a - t_*| < |b - t_*|, |a - b| < \max\{|a - t_*|, |b - t_*|\} \Rightarrow I(b) \leq I(a) \tag{5}$$

I define the ideal range of values as the set of values such that the individual function of any one of those values is the same as the individual function of the ideal tip. Thus, the ideal range is the set of inputs that maximize the value of the individual function. Similarly, I define the social range for a given social norm as the set of values such that the social function of any one of those values equals the social function of the social norm. Properties (4) and (5) imply that the social and ideal ranges respectively are

[51] When the social and ideal ranges overlap, it may be that t_k is not well-defined. In such a case, one can say that the consumer selects a value to tip at random among the values that maximize utility given the social norm. The exact rule for choosing t_k does not matter as long as t_k is a value that maximizes utility in period k.

connected, i.e. both the ideal range and the social range are intervals. These intervals must also be closed.[52]

Finally, it is necessary to specify certain properties about the individual and social functions. The purpose of this setup is to better define the minimal conditions necessary to have the social norm converge to the ideal range. While differentiability of the two functions is not necessary, one must assume that the partial derivative of the social function with respect to the individual tip evaluated at the social norm exists, and is zero:

$$\frac{\partial S}{\partial t_k}(n_k, n_k) = 0 \tag{6}$$

This is the condition that the social norm is not sharply defined. Graphically, equation (6) implies that there is no kink in the social function at the social norm. Intuitively, equation (6) says that the social pressure goes away at the norm, and is arbitrarily small as one gets arbitrarily close to the social norm. This makes sense; as long as one is close enough to the social norm, there should be an insignificant amount of social pressure to conform, as it is no longer clear if one is in violation of the norm.

I do not need the individual function to be differentiable anywhere, but I do need it to have a condition necessary for differentiability.[53] Specifically, assume the individual function to be strictly increasing as the tip approaches the ideal tip while outside the ideal range; furthermore, it does so in a manner that imitates having a nonzero derivative in the process. For any value x_* outside of the ideal range, we say that there must be some sequence x_k that converges to x_* such that there is some positive number *L* such that for all x_k:

$$|I(x_k) - I(x_*)| \geq L|x_k - x_*| \tag{7}$$

This equation simply says that the individual function changes in a rather persistent matter; it should not be the case that there is a range of values outside of the ideal range in which the individual is relatively

[52] See appendix A4 for a proof.
[53] A brief proof of necessity of condition in appendix A4.

unconcerned with the difference in how much s/he tips. It would simply not make sense if, for example, a customer was personally quite concerned with the difference between tipping one dollar and two dollars, relatively unconcerned with the difference between tipping two dollars and three dollars, and again concerned with the difference between tipping three dollars and four dollars.[54]

What results is the following theorem:

Theorem 1: *Given any initial social norm n_0, the social norm will approach some value within the ideal range over time.*

From this theorem, we can derive some corollaries:

Corollary: *If the ideal range is the ideal tip, then the social norm approaches the ideal tip over time.*

Corollary: *If the ideal tip is negative and the ideal range does not include or border zero, then the social norm for tipping will become negative in finite time.*

Of course, the strength of a social norm is certainly related to the rate at which the norm decays. To capture something of this effect, we express this final corollary:

Corollary: *If there is some real value ε such that for any social norm n_k the social range for that norm includes the interval $(n_k - \varepsilon, n_k + \varepsilon)$, then the social norm enters the ideal range in finite time and never leaves the range.*

In other words, as long as the social range remains substantial, societal pressure will completely disappear in finite time. However, as shown in Azar (2004), this condition is not necessary for convergence to occur in finite time.

Subsection III: The Persistence of Well-Defined Social Norms

[54] We could have this assumption fall out of a conviction that people experience money (or the ability to use money to make purchases) as an input to a convex utility function. This belief has an intuitive appeal; you hear a hint of it whenever people claim that the wealthy do not enjoy their money as much as the poor would. In addition, this idea has a distinguished pedigree, finding itself in the crux of the argument in Lange (1937) for the benefits of socialism over capitalism!

In the previous subsection, I established that social norms that are not sharply defined should go away in finite time, assuming reasonable conditions. However, the same does not hold for if the norm is sharply defined; instead, the norm may very well persist indefinitely. For example, suppose:

$$I(t) = 2n - t \tag{8}$$

$$S(t, n) = -2|t - n| \tag{9}$$

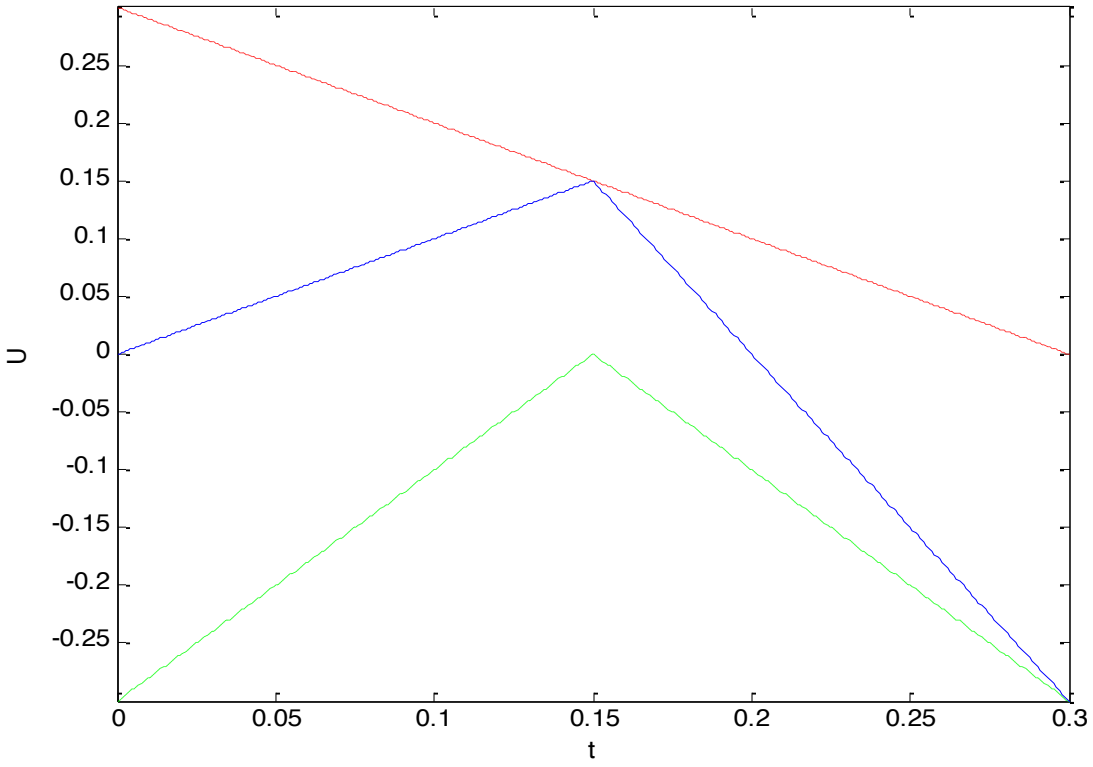
Note that I am momentarily suppressing the subscripts for simplicity. It's still obvious that the customer should not tip above the social norm. Yet the utility function for positive tips below or equal to the social norm can be expressed as:

$$U(t, n) = 2n - 2(n - t) - t = t \tag{10}$$

It's very obvious then that the utility function is maximized when $t = n$; the customer should tip the social norm exactly. This is described in the figure below.

But this is where the notion of the model customer breaks down. Until this point, it appeared reasonable to assume that everybody had identical utility functions, as it simplified the calculations while still leading to both powerful and generalizable results. However, in this case it is unreasonable to assume that everybody has a utility function such that they will conform to the norm every time, no matter how great the pressure to conform. At the minimum, there should be antisocial people who simply cannot be bothered to care about public opinion in the slightest. If there were rebels in puritanical New England and Victorian England, it seems unlikely that any society can completely suppress rebellion through societal pressure alone.

Figure A1: An Example of a Utility Function that Leads to Complete Conformity to the Norm



This is a graphical representation of functions (8), (9), and (10), with the social norm set to n=.15. The individual function is represented by the dotted red line, the social function by the dotted green line, and the utility function by the solid blue line. Note that both the social and utility functions are maximized at the social norm, whereas the individual function is never maximized everywhere, yet declines too slowly to matter.

Imagine instead that some people have utility functions as described above, in which there is a persistent drive towards conformity, while others have utility functions that always lead to a tip less than the social norm would indicate. We can even imagine some people whose utility functions indicate that they conform to the norm sometimes, and defy the norm at other times. It should be the case that the average amount tipped, and thus the social norm for tipping, should decline over time as the nonconformists continue to tip below the norm while others tip at the norm. Furthermore, it seems

pretty clear that the average tip should approach zero over time if there are a number of people each period who refuse to tip at all.

Theorem 2: *If there is some positive percentage of nonconformists that will always exist in a society and this positive percentage never tips above a certain value, then the social norm will approach or end up below that value.*

Corollary: If there is some positive percentage of nonconformists that will always exist in a society that will not tip at all, then the social norm will approach zero.

Of course, this would only play out if those who defy the norm are allowed to remain effective members of this society. There may be powerful motivations for members of this society to signal that those who deviate from the norm are no longer to be considered when considering what constitutes normal behavior. One can easily imagine such a prosecutor role being played by politicians or religious leaders, but it could also be played by protective parents or even young adults trying to prove their worth in this society.

It could be that everybody seeks to reaffirm to each other and defend their mutually held values, and thus act in concert to ostracize those who defy the norms established by society. However, it could also be the case that a norm has positive effects for societal welfare, and thus people feel obligated to defend this norm by blacklisting those who would threaten the established social order through rebellion against the social norm. Without going into much detail, it should be apparent that a society could take measures to prevent exposure to nonconformity; taking such precautions makes the survival of the norm a problem of whether or not the population of our society declines due to the excessive removal of its own members.

Such measures would be ineffective in the case of a soft social norm, as theoretically everybody would be defying the norm to some degree. Of course, some would be defying the norm more than others, and some people will undoubtedly surpass the norm. However, the only way to ensure

sufficient compliance would be to eliminate everybody from society who tipped below the previous average (in theory, everybody). Such a society would quickly become depopulated, and then effectively cease to exist. The point is that attempting to enforce a soft social norm such that it remains stable over time is clearly a ridiculous exercise, whereas it is quite believable that a society both could and would use some means of ostracizing the noncompliant in order to preserve a social norm indefinitely. This method need not even act as a deterrent as long as population decline isn't a significant issue.

However, it may be that a society has a slowly decaying social norm that they see no need to defend because the norm decays at an imperceptible rate. Indeed, this may very well be the case under many conditions, as it should be the case that the norm never truly goes away with time. Let us say that some positive percentage of the society will always be conformists. This would not be possible in the model laid out in the previous subsection, as the lack of sharpness of the social norm should have prevented the existence of conformists entirely. But allowing for the existence of these conformists results in the following theorem:

Theorem 3: If there is some positive percentage of nonconformists that will always exist in a society, and these nonconformists never tip, then the norm for tipping will approach zero over time. However, if there is some positive percentage of conformists that will always exist in a society, and the initial social norm for tipping is positive, then the norm for tipping will never reach zero; in any given moment, the norm for tipping will remain positive.

This theorem essentially makes clear that, under the reasonable assumption that there will always be a certain percentage of people who do not tip at all, and another percentage of people who tip as indicated by the norm of that time period, then the norm of tipping will decline towards zero over time, becoming closer and closer to zero as time goes on. Yet, though the norm may get very close to zero, it should never go away.

Now one may well argue that this is a silly claim to make: certainly the norm of tipping would be done away with once the norm reaches an absurdly low value. After all, why would people even bother to tip if the expected tip were only .01 percent of the bill? In fact, it seems hard to imagine that such a low tip wouldn't be perceived by the servers as insulting, even if such a tip is the norm. It's difficult to even imagine such a situation, for it appears absurd on its face. It seems inevitable that there is a threshold such that, once the norm dips below that threshold, the activity of tipping ceases altogether. The norm of tipping would have hit a sticky zero, and would thus cease to exist in any kind of meaningful manner.

However, although we know that such a threshold would be crossed in finite time, how long it takes to reach that threshold cannot be predetermined. In fact, as the norm approached the threshold, it makes sense to suppose that the rate at which the threshold is approached would decrease, at least in an absolute sense. Especially considering if there could be any short-run perturbations in this pattern caused by external events, or if the data collected on compliance with the norm is imperfect, it is easy to imagine that it could appear at times as if the norm is holding steady, or even growing for periods of time. If there are sufficiently few nonconformists, the norm would approach zero quite slowly, and thus it would be difficult to observe the deterioration of the social norm over limited periods of time.

Section IV: Discussion

These results obviously apply to almost any social norm with measurable economic implications. To bring up an example from before, monogamy can be perceived as a sharp social norm. Furthermore, the social pressure associated with monogamy in western society tends to be relatively strong, whereas there are relatively few nonconformists to be found. Yet undoubtedly there will always be at least a certain fraction of the population that practices polygamy (say .0001 percent). Thus, if one considers the social norm of monogamy to be primarily driven by compliance to a social norm, then one could make a parallel argument to the one above that our society will eventually see the day where

monogamy is no longer an effective social norm for our society. Furthermore, if the rate of decline of the social norm were to persist at the same rate that one may infer it to be declining at right now, and noting that the ostensible decline is too imperceptible to notice through casual observation, it would be the case that those around at the time when monogamy vanishes completely may not even notice at the time.

This is not to claim that the social institution of marriage is doomed. After all, that compliance with the social norm of monogamy is due to conformity alone is not only unproven, but unlikely. Still, this model suggests the ability to theoretically derive the slow deterioration of a social norm that appears to remain perfectly steady. To be able to make such conclusions about present-day norms in any sort of definitive manner seems ambitious for the moment. However, it certainly seems plausible that one may elaborate on this model to explore the slow decline of social norms in history, particularly when it appears that the norms take hundreds or even thousands of years to dissipate.

For example, although there is much more behind racism than mere conformity, there may be certain manifestations of racism correspond nicely with the notion of a sharply defined social norm with relatively few nonconformists, as well as some ability to punish nonconformists. In particular, social clubs that maintained a “whites only” policy could recognize when their members violated the norm by attempting to bring in outsiders, and even punish those who tried. Indeed, anecdotal evidence seems to indicate that the decline of these policies is strongly related to the proliferation of more tolerant notions of racial equality. This suggests that attitudes within these establishments changed because people’s personal attitudes changed (or they felt external pressure to change); in other words, these social clubs became filled with nonconformists, and the social pressure and norm-enforcement mechanisms could

no longer resist the changing times. Without the changed environment, it's difficult to imagine that the segregationist norm in various social clubs would have eroded much at all in the past few decades.[55]

This model is clearly too simplistic to be directly applied to real-world examples in a truly rigorous manner. But it does suggest new ways to consider the development of social norms over time. Furthermore, the model lays out a compelling argument for why conformity should not propagate most social norms indefinitely, while suggesting that there may be a few exceptions.

[55] Of course, there are almost certainly some social clubs that still hold a "whites only" policy. The above model suggests that these clubs will likely continue their exclusionist policies until changing societal values in the communities around these clubs create enough nonconformists within these clubs to affect change. Do not count on their norm of not admitting minorities to erode over time by itself.

Appendix A4: Proofs Related to A3

Proof of the Social and Ideal Ranges being Closed:[56]

There are several ways of proving it. One way is to rely on Theorem 1, although this may not be very convincing as it may appear that Theorem 1 relies on the ideal range being closed, and thus that such logic is circular. Instead, consider any continuous function *f(x)* and any connected set of values such that for any such value *y*, *f(y)* is constant. Let *z* be on the boundary of such a set. Then there is a sequence of values $\{y_k\}$ within the connected set that converge to *z*, implying that $f(y) = f(z)$, and thus that *z* is a member of the set in question. ■

Proof of Condition as Necessary for Differentiability with Nonzero Derivative:

Let *f(x)* be a function differentiable at point x_0 such that:

$$\frac{df}{dx}(x_0) \neq 0$$

Let *L* and ϵ be positive values such that:

$$L + \epsilon = |\frac{df}{dx}(x_0)|$$

Then there is some sequence of values $\{x_k\}$ such that for every value of *k*:

$$L + \epsilon - \frac{|f(x_k) - f(x_0)|}{|x_k - x_0|} < \epsilon$$

$$|f(x_k) - f(x_0)| \geq L|x_k - x_0| \quad ■$$

Proof of Theorem 1:

Let *L* be a positive number. Let $\{n_k\}$ be a sequence of social norms outside the idea range. For any value of *k*, there is some point x_k between n_k and t_* and sufficiently close to the social norm such that equations (6) and (7) imply that:

$$|I(x_k) - I(n_k)| - |S(x_k) - S(n_k)| > L(|x_k - n_k| - |x_k - n_k|) = 0 \quad (8)$$

$$|I(x_k) - I(n_k)| > |S(x_k) - S(n_k)| \quad (9)$$

[56] This is actually a more general proof that the connected preimage of a constant continuous mapping is closed.

By the monotonicity of the individual and social functions as previously described:

$$I(x_k) + S(x_k, n_k) > I(n_k) + S(n_k, n_k) \tag{10}$$

$$U(x_k, n_k) > U(n_k, n_k) \tag{11}$$

Since the utility function is continuous conditional on the norm, it achieves a maximum value on the closed interval between n_k and t_* inclusively. Furthermore, since leaving this interval causes both the social and individual functions to decrease, the local maximum within this interval must also be the global maximum. The above conclusion shows that the maximum must not be attained at the social norm, as there is clearly a value for the tip for which the utility function is greater. The sequence $\{t_k\}$ must be converging because it is a bounded monotonic series of real values; let us say that there is some value n_* such that $n_k \to n_*$ as $k \to \infty$.

Suppose that n_* is not in the ideal range. Then the above result implies that there is some value within the interval $[t_*, n_*)$ that maximizes the utility function for the given social norm n_*.[57] Denote this value by u_*, so that $|u_* - n_*| > 0$. Note that the maximization function is continuous, so the composition of the maximization function and the utility function is also continuous. Furthermore, note that:

$$u_* = \text{argmax}_x\, U(x, n_*) \tag{12}$$

Thus, equation (12) is continuous.

Let $\epsilon = |u_* - n_*| > 0$. By continuity, $\exists \delta > 0$ such that $n \in (n_* - \delta, n_* + \delta)$ implies that $|u_* - \text{argmax}_x\, U(x, n)| < \epsilon$. But $\exists k$ such that $n_k \in (n_*, n_* + \delta)$, implying:

$$n_{k+1} = t_k = \underset{x}{\text{argmax}}\, U(x, n_k) \in [t_*, n_*)$$

This contradicts the definition of n_*. It must be the case that t_* is in the ideal range. ■

The first corollary is relatively obvious.

[57] Technically this assumes that $t_* < n_*$. This is some abuse of notation done for simplicity, as all that is really needed is to indicate connected intervals and what of the interval's boundary is included in the interval. Informally speaking, one may ignore the issue of whether the ideal tip is less than the social norm; formally, we assume without loss of generality that this is the case.

Proof of second corollary:

Choose n_0. Let n_* be the point to which the series $\{t_k\}$ converges. Note that Theorem 1 implies that this point must be within the ideal range, and thus must be negative. From the definition of convergence, there is a positive integer *K* such that $\forall k \geq K$:

$$|n_* - t_k| < |n_*| \tag{14}$$

Since n_* is negative, it follows that so is t_k. So for any social norm there is some positive integer *K* such that the social norm for tipping is negative in *K+1* periods of time and remains negative thereafter. ■

Proof of third corollary:

Choose n_0. Let n_* be the point to which the series $\{n_k\}$ converges. There is a positive integer *k* such that $|n_* - n_k| < \varepsilon$. Note that the social function is constant between n_k and n_*, while the individual function increases as the input approaches the ideal tip until the input reaches the boundary of the ideal range. So any value between n_k and n_* that is within the ideal range is a value for the tip that maximizes the utility function. Thus, the tip for period *k+1*, and thus the social norm for period *k+2*, is within the ideal range.

For period *k+2*, the norm is within the ideal range, so any tip within the intersection of the social and ideal ranges will maximize the utility function. In fact, for any integer $l > 1$, n_{k+l} is within the ideal range, implying that n_{k+l} is within the ideal range. By induction, the social norm never leaves the ideal range. ■

For the proofs in A3, I need the following lemma:

Lemma: Given $a, b \in [0,1]$ *such that* $a + b = 1$ *and letting* $t_{n+1} = at_n + bt^*$ *for* $\{t_n\} \subset \mathbb{R}$ *and* $t^* \in \mathbb{R}$*, it follows that the sequence* $\{t_n\}$ *will converge somewhere between* t_0 *and* t^*.

Proof: First, note that if the sequence $\{t_n\}$ will converge, then it will converge between t_0 and t^* by the Squeeze Theorem, as $\{t_n\}$ is bounded between those two values. Now observe that there are two possibilities. Either $a = 1$, in which case $\{t_n\} = \{t_0\}$ and the lemma is obvious, or $a \neq 1$, implying:

$$\forall n \in \mathbb{N}\ t_{n+1} \in [t^*, t_n)$$

The notation above assumes that $t^* < t_n$, but it is obvious that this is an assumption only of notational simplicity. But from here, it is clear that the sequence $\{t_n\}$ is monotonic and bounded. Since this is a sequence of real numbers, the sequence must converge. ■

Proof of Theorem 2:

While conceptually important, this proof is relatively simple. Let t_n denote the average amount tipped in period *n*. Let t^* denote the value above which all of the nonconformists refuse to tip. Suppose that some positive fraction of the population, $a \in (0,1]$, will always tip at or below the value t^*. Of course, no one will tip more than the average amount tipped last time. So for any $n \in \mathbb{N}$:

$$t_{n+1} \leq (1-a)t_n + at^*$$

Solving and induction yields:

$$\forall N \in \mathbb{N}, t_N - t^* \leq (1-a)(t_{N-1} - t^*) \leq \cdots \leq (1-a)^N(t_0 - t^*)$$

But then:

$$\lim_{n\to\infty} t_n - t^* \leq t_0 \lim_{n\to\infty} (1-a)^n(t_0 - t^*) = 0$$

Thus, $\lim_{n\to\infty} t_n \leq t^*$. This proves the theorem. ■

For the corollary, set $t^* = 0$ *from the above proof. Note that* $\lim_{n\to\infty} t_n \not< 0$*, so* $\lim_{n\to\infty} t_n = 0$. ■

Proof of Theorem 3:

The first part of this theorem is the above corollary. However, let $b \in (0,1]$ be the fraction of people who will always tip the social norm. Furthermore, the theorem assumes that $t_0 > 0$. Since nobody can tip a negative amount, for any $n \in \mathbb{N}$ induction demonstrates:

$$t_n \geq bt_{n-1} \geq \cdots \geq b^n t_0 > 0$$ ■

Appendix A5: Nash Equilibrium Convergence

A *Nash Equilibrium* is a set of choices for the players of a game such that no player wishes to change his/her choice given the choices of the other players. Let $\left(x_{1,1}^n, \dots, x_{1,k_1}^n, x_{2,1}^n, \dots, x_{2,k_2}^n, \dots, x_{N,k_N}^n\right)$ be an ordered set of real values, where $x_{j,l}^n$ is the lth choice of the jth player, for $j \in \{1, \dots, N\}$ and $l \in \{1, \dots, k_j\}$, while n represents the period, so that $n \in \mathbb{N}$. Let $U_j\left(x_{1,1}^n, \dots, x_{1,k_1}^n, x_{2,1}^n, \dots, x_{2,k_2}^n, \dots, x_{N,k_N}^n\right)$ be the utility function for the jth player. Let $\left(S_{1,1}^n, \dots, S_{1,k_1}^n, S_{2,1}^n, \dots, S_{2,k_2}^n, \dots, S_{N,k_N}^n\right)$ represent the Nash equilibrium, i.e., every utility function is maximized for a player given the choices of the other players. In other words, for any $x_{j,1}^n, \dots, x_{j,k_2}^n \in \mathbb{R}$:

$$U_j\left(S_{1,1}^n, \dots, S_{1,k_1}^n, S_{2,1}^n, \dots, S_{2,k_2}^n, \dots, S_{N,k_N}^n\right) \geq U_j\left(S_{1,1}^n, \dots, S_{1,k_1}^n, S_{2,1}^n, \dots, S_{2,k_2}^n, \dots, x_{j,1}^n, \dots, x_{j,k_2}^n, \dots, S_{N,k_N}^n\right)$$

The following theorem is the strongest mathematical result; the corollary that follows is more relevant to this paper:

Theorem: For every convergent subsequence of $\left(S_{1,1}^n, \dots, S_{1,k_1}^n, S_{2,1}^n, \dots, S_{2,k_2}^n, \dots, S_{N,k_N}^n\right)$ that converges to a point where the utility function of every player is continuous, then the point where the subsequence converges is a Nash Equilibrium for these functions when choices for each player are restricted to the set of points on which the utility function of that player is continuous.

Proof: This is a proof of the contrapositive. Namely, if there is a point $\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right)$ such that there is some $j \in \{1, \dots, N\}$ such that U_j is continuous at $\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right)$ & $\left(S_{1,1}, \dots, S_{N,k_N}\right)$, and $U_j\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right) > U_j\left(S_{1,1}, \dots, S_{N,k_N}\right)$, then $\left(S_{1,1}^n, \dots, S_{1,k_1}^n, \dots, S_{N,k_N}^n\right)$ does not converge to $\left(S_{1,1}, \dots, S_{N,k_N}\right)$. To see this, let $\epsilon = U_j\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right) - U_j\left(S_{1,1}, \dots, S_{N,k_N}\right) > 0$. Then there are two positive numbers δ_S and δ_x such that when some vector $\left(t_{1,1}, \dots, t_{N,k_N}\right)$ of the inputs into U_j get within those numbers of $\left(S_{1,1}, \dots, S_{N,k_N}\right)$ or $\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right)$, respectively, then $U_j\left(t_{1,1}, \dots, t_{N,k_N}\right)$ is

within $\frac{\epsilon}{2}$ of U_j of $\left(S_{1,1}, \dots, S_{N,k_N}\right)$ or $\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right)$, respectively. $\exists\, n$ sufficiently large such that $\left\|\left(S_{1,1}, \dots, S_{N,k_N}\right) - \left(S_{1,1}^n, \dots, S_{N,k_N}^n\right)\right\| < \frac{\epsilon}{2}$. Then I can note the following:

$$U_j\left(S_{1,1}^n, \dots, x_{j,1}^n, \dots, x_{j,k_j}^n, \dots, S_{N,k_N}^n\right) > U_j\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right) - \frac{\epsilon}{2}$$

$$U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right) < U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right) + \frac{\epsilon}{2}$$

Plugging in for ϵ yields:

$$U_j\left(S_{1,1}^n, \dots, x_{j,1}^n, \dots, x_{j,k_j}^n, \dots, S_{N,k_N}^n\right) > \frac{U_j\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right) - U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right)}{2}$$

$$U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right) < \frac{U_j\left(S_{1,1}, \dots, x_{j,1}, \dots, x_{j,k_j}, \dots, S_{N,k_N}\right) - U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right)}{2}$$

Thus, $U_j\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right) < U_j\left(S_{1,1}^n, \dots, x_{j,1}^n, \dots, x_{j,k_j}^n, \dots, S_{N,k_N}^n\right)$; $\left(S_{1,1}^n, \dots, S_{N,k_N}^n\right)$ is not a Nash Equilibrium. If a point is not a Nash Equilibrium, then there is not a convergent subsequence of Nash Equilibria that converges to that point. ■

Corollary: If all utility functions are continuous on their domains, then all cluster points of sequences of Nash Equilibria are Nash Equilibria.

Proof: Cluster points are limits of subsequences of a sequence. Also, the utility functions are continuous at the cluster points. The theorem then implies the corollary. ■

Appendix A6: How the BPDP Model Explains Irrational Cooperation in the Prisoner's Dilemma

Let there be two players, player 1 and player 2, who are playing N identical rounds of the prisoner's dilemma game (where $N \in \mathbb{N}$). Each player has the same utility function as the other, and each player's present utility is standardized to zero in a round when both players cooperate. The benefit to a player's present utility for betraying the other player is B, while the amount that a player's present utility is hurt by the betrayal of the other player is S. So the present utility that a player gets from both players betraying each other in a round is $B - S$. The discount rate from one round to another is $\theta \in [0,1]$. The classical result for such a game played by rational actors is that both players betray each other immediately and for every trial, as in Roth and Murnighan (1978).

However, suppose that both players got to choose to pre-commit to the strategy of cooperating in the first round and continuing to cooperate until the other player chose to betray the first player. If a player chooses this strategy, that player sends a signal to the other player. The player only enacts this strategy if s/he receives the same signal from the other player. However, there is a probability $1 - p$ (with $p \in [0,1]$) that the player can send this signal even without pre-committing to this strategy. Both players know whether or not this strategy is available to them, and also the probability (which is the same for both players) that the other player can send a false signal. I assume that a player will always send this signal without pre-committing if possible, since this way the player gets the possible benefit of the signal along with the freedom of acting in a rationally self-interested manner.[58]

For the purpose of simplicity, consider only pure strategy equilibria. Then the following result holds:

Theorem: Let both players send a signal of cooperation, but player 1 sends it falsely. If $B > p\theta S$, then the only Nash equilibrium is for player 1 to betray player 2 in the first round, and likewise for all rounds

[58] This is technically not a weakly-dominating strategy, since it is possible that sending the signal is no better than not sending the signal. For example, if $p = 1$, then the signal is meaningless, and doubtlessly the two players will immediately and consistently betray each other, as if the players had not sent the signal at all.

thereafter. If $B < p\theta S$, then the only Nash equilibrium is for player 1 to cooperate until either the last round or when player 2 betrays player 1, at which point both players will subsequently betray each other.[59]

Proof: To prove this result, I will prove the following statement by strong induction:

For $j \in \mathbb{N} \cap [2, N]$, it is a Nash equilibrium to betray the other player if $B > p\theta S$, and is not if $B < p\theta S$ on the jth last round.

First, examine the case where $j = 2$. If one of the players has already betrayed the other, than both players play like rational actors, and the Nash equilibrium is for both players to betray each other. Otherwise, each player thinks that the other player may be committed to the strategy of rewarding cooperation and punishing defection; the probability that this is the case is p. So if one player betrayed the other, the player would gain B in utility, but would lose S in utility the next round if the other player is playing the retaliatory strategy. So in that case, the net change in utility would be $B - \theta S$. But if player 2 is sending a fake signal of pre-commitment, then player 2 will doubtlessly choose betrayal in the last round, since there is no prospect for punishment in subsequent rounds. Thus, the net benefit of player 1 betraying player 2 if the player 2 is sending a fake signal is B. Thus, the expected change in utility from player 1 choosing betrayal over cooperation in the second last round is $B - p\theta S$. The statement follows.

Now, assume that player 1 will betray player 2 in the j-1st last round and all subsequent rounds. Then again, by the same logic as above, the net change in expected utility from choosing betrayal over cooperation in the jth last round is $B - p\theta S$. This proves the statement.

Since the statement holds for all such $j \in \mathbb{N} \cap [2, N]$, it will be worthwhile to consistently betray player 2 if $B > p\theta S$. However, if $B < p\theta S$, then player 1 will choose to cooperate with player 2 until

[59] Note that the trivial case of $B = p\theta S$ is ignored throughout this exposition. In such a situation, player 1 is indifferent between betraying or cooperating with player 2 at any round in which neither player has betrayed the other yet, with the exception of the last round.

player 2 betrays player 1, or until $j = 1$, at which point there will be one round remaining, and it will be in the interest of player 1 to betray player 2. ■

It readily follows from this result that if both players signal falsely, and if $B < p\theta S$, then both players will cooperate until the other initiates the cycle of betrayals or the last round arises. But since both players are waiting for the other to betray first, neither player will betray the other until the last round, at which point both players will betray each other.

Now suppose that player 1 cannot falsely signal the cooperative pre-commitment strategy. If player 1 fails to pre-commit to this strategy, then both players will betray each other every round, and player 1 will have net present discounted utility $(B - S)\frac{1-\theta^N}{1-\theta}$.[60] Player 1 will have the same present discounted utility upon pre-committing if player 2 does not send the pre-commitment signal as well. However, if player 2 also sends the pre-commitment signal, then there is a probability p that the signal is genuine, in which case the net present discounted utility will be zero, and a probability $1 - p$ that the signal is false, in which case the result depends on the size of p. Specifically, let $FA(p)$ be the present discounted utility of player 1 pre-committing if player 2 falsely signals pre-commitment. Then:[61]

$$FA(p) = \begin{cases} -\theta^{N-1}S, & B < p\theta S \\ B\dfrac{\theta - \theta^N}{1-\theta} - S\dfrac{1-\theta^N}{1-\theta}, & B > p\theta S \end{cases}$$

So if $B < p\theta S$, then the expected present discounted utility for player 1 to pre-commit to the cooperative strategy if player 2 also signals pre-commitment is $(p-1)\theta^{N-1}S$. If $B > p\theta S$, then the expected present discounted utility is $(1-p)\left[B\frac{\theta-\theta^N}{1-\theta} - S\frac{1-\theta^N}{1-\theta}\right]$. From these statements, we get the following result:

[60] If $\theta = 1$, then the net present discounted utility is $(B - S)N$.
[61] If $\theta = 1$, then FA is either $-S$ or $B(N-1) - SN$. Also, note that I am still ignoring the trivial case of $B = p\theta S$.

Theorem: If $B > \theta S$, or if p is sufficiently close to zero, then the players will not bother to pre-commit to cooperation. However, if $B < \theta S$, then if p is sufficiently close to unity, players will signal pre-commitment.

Proof: The first part of this theory is readily clear. If $B > \theta S$, then neither player can reward the other sufficiently to compensate each other for not betraying each other. So choosing not to pre-commit weakly dominates choosing to pre-commit. Similarly, for a sufficiently small p, $B > p\theta S$, and so a player who chose to pre-commit would be nearly assured to merely lose out on an opportunity to betray the other player in the first round. Formally, the limit as p approaches zero of the expected present discounted utility of a player choosing to pre-commit to cooperation and receiving the signal from the other player is $B\frac{\theta-\theta^N}{1-\theta} - S\frac{1-\theta^N}{1-\theta}$. For no commitment, it is $(B-S)\frac{1-\theta^N}{1-\theta}$, so it is clearly in the interest of a player to avoid the cooperation strategy when p is sufficiently close to zero.

Regarding the second part of the theorem, if a player can send a false signal, then s/he will do so by assumption. Otherwise, if $B < \theta S$, then one can choose p sufficiently close to unity such that $B < p\theta S$, in which case the expected present discounted utility of a player choosing to pre-commit and receiving a pre-commitment signal from the other player is $(p-1)\theta^{N-1}S$, a value which becomes arbitrarily close to zero as p approaches unity. Thus, for a value of p sufficiently close to unity, the expected present discounted utility of choosing to cooperate and getting the signal of cooperation from the other player becomes very close to zero. Since $B < \theta S \Rightarrow B < S$, the present discounted utility of not pre-committing to cooperation is negative, implying that for a value of p sufficiently close to unity, the strategy of pre-committing to cooperation dominates the strategy of refusing to pre-commit.

■

Note that it is reasonable to take the limit as N approaches infinity in order to approximate expected utility calculations when there are a large number of trials. In such a situation, if $\theta < 1$, then the utility of failing to signal cooperation is $\frac{B-S}{1-\theta}$ and the expected utility of committing to cooperate with

the other player is either zero if $B < p\theta S$ or $(1-p)\frac{\theta B - S}{1-\theta}$ if $B > p\theta S$. Note that it will be worthwhile for player 1 to attempt to cooperate with player 2 if $B < p\theta S$, but it is unclear whether or not it will be if $B > p\theta S$.

In conclusion, if the benefits of betrayal do not outweigh the product of the probability of the other player being able to falsely signal, the discount rate per round, and the disutility of being betrayed by the other player, then situations can arise in which both players choose to cooperate, and do not betray each other at all. Furthermore, since $B > p\theta S \Rightarrow B > S$, it is always better for 'society', or both players combined, that this situation arises. So if a society has to deal with repeated prisoner's dilemmas, it may be beneficial to make p as close to unity as possible, while also having trials as frequently as possible. The ideal situation for a society is for $p = \theta = 1$, in which case there will be cooperation for all trials when $B < S$, and betrayal for all trials when $B > S$. This situation can be approximated when trials are repeated quickly and those who are untrustworthy tend to develop a bad reputation and become ostracized.

However, the simplicity of this model should be emphasized. The discount rate and probability of a false signal remain constant for all trials, an assumption that may not be realistic. By adjusting these values, it may be possible to get players to switch from cooperation to defection in the middle of the game. Furthermore, when using these values as constants, it is possible to also show that a rational actor will react identically to the tit-for-tat strategy as s/he will to the strategy described above. In the tit-for-tat strategy described in Axelrod and Hamilton (1981), a player cooperates in the first round, and then mimics the behavior of the other player in the kth round in the k+1st round thereafter (for k=2,3,…). So a rational actor playing against the tit-for-tat strategy will always cooperate if $B < p\theta S$, except in the last round when betrayal is inevitable, and will always betray if $B > p\theta S$. Furthermore, two players that are both using the tit-for-tat strategy will cooperate the whole time, as will one player using the tit-for-tat strategy who plays with a player using the strategy introduced in the beginning of this section.

However, the interchangeability of these two strategies may not hold if p and θ are not treated as constants.

Note that Roth and Murnighan (1978) show that the existence of cooperative Nash equilibria results in a statistically significant increase in the frequency of cooperation in the prisoner's dilemma. However, they were looking at iterative games in which there was a geometric distribution of trials rather than a fixed number of trials. The iterative games that they looked at are similar to this version of the iterated prisoner's dilemma only when there are a guaranteed infinite number of trials.

Yet it should be possible for the BDPD model to be extended to look at the data of Roth and Murnighan (1978). In particular, it is of immediate interest to attempt to explain why the existence of cooperative Nash equilibria with rational actors would lead to increased cooperation in the BDPD model. In the process, one would need to determine how actors in the BDPD model choose between multiple Nash equilibria. Perhaps players could signal to each other the equilibrium that they will choose, and restrictions on their ability to signal to each other would lead to each selecting the wrong equilibrium.

Appendix A7: A Few Notes on How One Might Formally Model the Development of Social Norms

In this paper, I have avoided modeling the general development of social norms due to the complexity of the process. Members of society signal to each other, and the concepts developed by society help to interpret those signals. In addition, people are not necessarily maximizing utility when they signal; for any particular individual, the planner may have already pre-committed them to certain social norms, thus restricting the "possible" utility-maximizing behavior. Furthermore, various situations can arise in which different social norms may be seen to conflict with each other.

However, there are readily available means by which one can model special cases of this general process. If one were to model the development of social norms in a society that had no other social norms, then one could merely describe the behavior of the individual members of society as rational actors. One could then employ graph theory as in standard game theoretic analysis. Such would be a general case of the work done in the previous section.

But usually people already have pre-commitments to other social norms. Yet one could assume that it was always perfectly clear when social norms conflicted with each other. Then one could assume that nobody would pre-commit to a social norm that contradicts a social norm to which they have already pre-committed. If one then assumes that there is a certain rate at which members of society are replaced with members who have not pre-committed to any social norm, one can readily model the development of social norms in such a society while still employing the standard tools of game theory.

Yet this is a critical simplification from how reality works. It is frequently unclear when social norms contradict each other, and there are a plethora of cases in which people have embodied social norms that almost surely contradict each other in situations that arise quite often. People make promises to multiple friends, only to realize that they cannot keep one promise without violating another. One may have pre-committed to avoid both telling a lie and hurting a friend's feelings, which creates a difficult situation when one's friend asks for an opinion on a subject matter where a truthful

response would hurt the friend's feelings. In that situation, even refusing to respond would indicate that the response would not be to the friend's liking, which could also hurt the friend's feelings.

To completely model cases such as these, one must formally introduce cognitive dissonance. There are a couple of ways in which one could do this. The simpler method would be to rework the pre-commitment mechanism as implying a cost to the actor for violating the social norm. Then, the problem becomes reduced to one of maximizing utility. If the doer has no personal stake in the decision, then the norm with the lower cost of being violated will be violated. If one wished, one could then specify that the actor was no longer pre-committed to the violated norm, or that the activity performed is no longer considered to be in violation of the norm.

However, this method has a couple of drawbacks. One, if one has a norm no longer apply to a situation in which it had been violated, it is not clear whether or not the norm should apply to a similar situation. Two, this method of modeling the decision between two norms does not take into account the phenomenon of being unsure of whether a norm applies.

Alternatively, one might employ fuzzy logic in order to showcase the process of cognitive dissonance. One could represent a pre-commitment to a social norm with a distribution. There may be a range such that integrating the distribution over any point within that range is one.[62] The implication is that any point within that range is a point at which it is clear that the norm would be violated. The distribution could also never be below zero.[63] In any given possible scenario, one could integrate the distribution over the relevant set of choices, and the resulting value could be plugged into a monotonically increasing cost function to yield the psychological cost to the actor of making those decisions. Formally, if C represents the set of choices, f represents the distribution, and $c: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ represents the cost function, then:

[62] An example of this is the delta distribution, which evaluates to one when a particular point is within the range of integration, and evaluates to zero otherwise.

[63] One could plausibly have a norm that rewards extra compliance as well as punishing violations. In that case, a negative distribution would make sense.

$$c(0) = 0$$

If a norm is not violated at all, there should be no psychological cost. Furthermore, $c\left(\int_C f\right)$ represents the psychological cost. That is to say, the psychological cost is the cost function evaluated at the integrated magnitude of the violation of the norm.

However, cognitive dissonance involves a process of resolving the dissonance through changing one's beliefs. One could embody this process by multiplying the distribution by another nonnegative distribution. This maintains the intuition that a social norm cannot arise out of the process of cognitive dissonance. However, one that already exists could become stronger through this process. In addition, a social norm that is continuously violated could cease to exist even if it is not in conflict with another social norm. Thus, through the process of cognitive dissonance, one need not replace members of society with others without pre-commitments in order to let social norms die out.

Appendix A8: Program in Matlab Used to Get Voting Probability Results

Note that the following program only works for even N.

```
function voting(p,N)
%Calculates the odds of being a pivotal voter given that each voter
has
%probability p of voting for candidate A, 1-p for voting for candidate
B,
%the probability of any voter voting for a candidate is independent of
the
%votes of everyone else, and N people vote.
%   For each j, figures out the probability of being a pivotal voter
%   given 2j people are voting.  Call this probability d.  When j=0,
%   d=1.  For positive j, d(2j)=2*p*(1-p)*d(2j-2)*(2*j-1)/j.  Keep N
%   even for now.

d = 1;
for j=1:(N/2)
    d = 2*p*(1-p)*d*(2*j-1)/j;
end
d
end
```